

LendingClub Loans Pricing

HarvardX - PH125.9x Data Science Capstone

Emmanuel Rialland - https://github.com/Emmanuel_R8

November 14, 2019

Contents

Introduction	5
1 Internal Rate of Return, Credit Margins and Net Present Values	7
1.1 Important warning	7
1.2 Background	7
1.3 Internal Rate of Return	8
1.4 Dataset calculation	9
1.4.1 IRR	9
1.4.2 Credit Margins	10
1.4.3 NPV	10
2 Dataset	11
2.1 Preamble	11
2.2 General presentation	11
2.2.1 Business volume	11
2.2.2 Loan lifecycle and status	13
2.2.3 Loan application	14
2.3 Rates	14
2.3.1 IRR and required credit margins	14
2.3.2 Dataset	15
2.3.3 Interest rates	15
2.3.4 Purpose	17
2.3.5 Payments	19
2.4 Net present value	20
2.4.1 Average NPV and credit margin by subgrade	20
2.4.2 NPV distribution by rating	21
2.5 Variables	21
2.5.1 Identification	24
2.6 Loan decision	24
2.7 Payment-related information	25
3 Modelling	26
3.1 Introduction	26
3.1.1 Split	26
3.1.2 The modeling task	26
4 Stochastic Gradient Descent	27
4.1 Description of the model	27
4.2 Gradient descent	27
4.2.1 Generalities	27
4.2.2 Cost function as a function of the Q parameter	28

4.3	Stochastic Gradient Descent	29
4.3.1	Stochastic Gradient Descent (SGD)	31
5	Conclusion	36
Appendix		37
5.1	List of assumptions / limitations regarding the dataset	37
5.2	Data preparation and formatting	37
5.2.1	LendinClub dataset	38
5.2.2	Zip codes and FIPS codes	43
5.2.3	Market interest rates	43
5.2.4	Macro-economical data	45
5.3	List of variables	47
5.4	Calculations of the internal rate of returns and month-to-default	53
5.4.1	R code	53
5.4.2	Julia code	56
5.4.3	Maxima derivation of the cost function	65
5.5	System version	66

List of Tables

2.1	Number of loans per status	12
2.2	Matured loans per status	24
5.1	Description of the dataset variables as provided in the dataset downloaded from Kaggle	47

List of Figures

2.1	Business volume written per month	12
2.2	Credit margins per grade over time	15
2.3	Credit margins per grade over time	16
2.4	Credit margins per grade over time	16
2.5	Interest rates given rating	16
2.6	Interest rate per grade over time	17
2.7	Historical Swap Rates	18
2.8	Credit margins per grade over time	18
2.9	Histograms of credit margins per purpose	19
2.10	Boxplots of credit margins per purpose	20
2.11	Distribution of NPV (%) depending on rating	22
2.12	NPV % over 120% (no y-axis scaling)	22
2.13	NPV % around 41% (no y-axis scaling)	23
2.14	NPV % around -1% (no y-axis scaling)	23
2.15	NPV % for close to total loss % (no y-axis scaling)	24
2.16	Funding and Write-offs by Sub-grades	25
4.1	Scikit Learn algorithm cheat-sheet	28

Introduction

Lending Club (*LC*) is an American company listed on the New York stock exchange that provides a platform for peer-to-peer lending. Unlike banks, it does not take deposits and invest them. It is purely a matching system. Each loan is split into \$25 that multiple investors can invest in. LC is remunerated by fees received from both sides. LC states that they have intermediated more than \$50bln since they started operations. Further description of the company is easily available online numerous sources.

In order for investors to make the best investment decisions, LC make a historical dataset publicly available to all registered investors. This dataset is the subject of this report. It was downloaded from the Kaggle data science website¹.

The size of the dataset is rich enough that it could be used to answer many different questions. We decided for a focused approach. Following Chapter 5 of (Peng, 2012), we will first formulate the question we want to answer to guide our analysis.

The business model of LC is to match borrowers and investors. Naturally, more people want to receive money than part with it. An important limiting factor to LC's growth is the ability to attract investors, build a trusting relationship where, as a minimum first step, investors trust LC to provide accurate, transparent and reliable information of the borrowers. For this purpose, LC decided not only to provide extensive information about potential borrowers' profile, but also historical information about past borrowers' performance. This is, as we understand, one of the key purposes of this dataset. We decided to use the dataset for this very purpose. Essentially, the questions are: **given a borrower profile, is his/her rating appropriate in terms of risk of default? And if a default occurs, what is the expected recovery?** The summary question is: **given a borrower profile, is the risk/reward balance appropriate to commit funds?** In answering this question, we understand that LC allows investment of very granular amounts. Therefore, even an individual investor can diversify his/her loan and risk portfolio. It is not necessary to 'gamble' funds on a single borrower. This is exactly what institutional investors achieve through syndication (although on a very different scale, typically \$10-25mln for a medium-size bank).

For this exercise, we made two simplifying (hopefully not simplistic) assumptions:

- In determining the risk/return balance, we have not accounted for LC's cost of intermediation. By ignoring fees paid by both sides, we obviously overestimate the returns to the investors. But in first approximation, **we will assume that the risk/reward balance, from the investors' point of view, across ratings is independent from fees.** This is a simplification. Real-world fees are higher the lower the investment grade and push the investors to receive, and the borrowers to pay, higher interest margin.
- All-in interest rates paid by borrowers are fixed. This is highly desirable for borrowers to be able to manage their cashflow. However, an investor should always consider an investment

¹<https://www.kaggle.com/wendykan/lending-club-loan-data/data>

return as a margin above a risk-free return. Banks would look at LIBOR; bond investors (e.g. life insurers) would look at government bonds. Those risk-free rates can change very quickly, whereas we understand that LC sets those rates on a less frequent basis. In other word, the risk premium will vary rapidly. **We assume that individual investors are ‘in-elastic’ to change in implied risk premia.** But we recognise this as a limitation of our work.

This report is organised as follows:

- [XXXX]

Chapter 1

Internal Rate of Return, Credit Margins and Net Present Values

In this section, we describe the response variables that we will generate for each loan and that will be used in the rest of this report.

As indicated in the introduction, our purpose is to test a model that predicts the financial risk of a loan.

1.1 Important warning

The calculations presented here are simplistic, although they bear some resemblance to what financial institutions (*FIs*) do. The literature on credit assessment and pricing is very rich and very complex. Finding the optimal capital allocation to particular risks while at the same time satisfying internal risk policies and regulatory requirements is a problem that financial institutions have yet to solve in full. Investing in a loan is not only a matter of assessing the risk of a particular borrower, but also assessing systemic risks (which exist across all borrowers), risks associated with funding the loan (interest, currency and liquidity markets), each requiring a risk assessment and pricing.

In other words, nobody would, let alone should, make any investment decision based on the calculations below.

1.2 Background

This subsection can be skipped by anybody with basic financial knowledge.

A bird in hand or two in the bush; a penny today or a pound tomorrow. What is the price of delaying obtaining and owning something? This is what pricing a loan is about. A lender could keep his/her cash in hand, or lend it and have it in hand later. He/she would accept this in exchange for receiving a bit more: this is the rate of interest. A lender wants to be compensated for delaying the possibility of using the cash, but also for taking the risk of not receiving it, partially or in full, when repayment is due.

There are borrowers that one can see as (almost) completely safe or risk-free such as central banks or governments of strong economies. A lender always has the possibility to lend to them instead of more risky borrowers. Therefore, a lender would require a higher interest rate than risk-free. The additional interest that a lender requires is commensurate with the risk of the borrower not repaying (called *credit worthiness*) and is called the *credit margin*.

For each individual borrower, an FI would assess information provided by the borrower and massive amounts of historical data to answer the question: considering historical borrowers with a profile similar to the applicant's, what is the probability of not getting principal and interest back (*Probability of Default* or *PD*)? And, in case the borrower stops paying and using additional courses of action (such as seizing and selling assets), what is the total loss that could be expected on average (*Loss given Default* or *LGD*)?

Making that assessment, the FI would require an interest rate which would roughly be the sum of:

- the risk-free rate;
- a margin to cover the average loss of similar individual borrowers¹;
- a margin to cover all the operational costs of running their operations; and,
- a margin to remunerate the capital allocated by the FI (banking regulations require all banks to allocate an amount of capital against any risk taken; those are stipulated in a number of complex rules).

Said crudely, this total is the amount for the FI to get out of bed and look at a loan. Although this sounds like an exact science (for some definition of the word), it is not. At the end of the day, the FI will also have to contend with the competition from other FIs or non banking lenders, market liquidity (if there is a lot of money available to be lent, it brings prices down) and, critically, whether the borrower would at all be interested in accepting that cost.

Note that the dataset is distorted by this additional survival effect: the application information of many loans does not appear merely because the rate of interest was considered too high (this is not dissimilar to *survival effects* where some data did not survive through the history of a dataset²).

1.3 Internal Rate of Return

For the purpose of this report, we will simplify things enormously: we will only consider the first two components of the interest rate. The risk-free rate and the credit margin that would cover the cost of default/losses of individual borrowers.

The modeling exercise will focus on trying to approximate *on average* the credit margin given the information provided by the borrower.

With respect to a given loan and its cash flow, two calculations are important here: the Net Present Value (*NPV*) and the Internal Rate of Return (*IRR*). If we remember that an FI is indifferent to holding a principal P today or receiving it with an annual interest a year later (i.e. $P \times (1 + r)$ where r is the annual rate of interest), we can say that any amount CF_1 received in a year is equivalent to $CF_0 = \frac{CF_1}{1+r}$ today. More generally, a stream of future cash receipts is worth:

$$NPV(r) = \sum_{Year=1}^{Year=n} \frac{CF_i}{(1+r)^i}$$

¹It is important to realise that the average margin only brings the borrower back to having earned the risk-free rate average: the additional income from the credit margin will be spent to cover average losses. In addition, we present the credit margin as income against borrower-specific losses. It does not address a lot of other risks such as correlation risks: a borrower might default because the economy as a whole gets worse, in which case many borrowers will default. This is a cyclical *systemic* risk similar to the 2007 US real estate crisis.

²A well-known example is historical stock prices which disappear when companies are de-listed or go bankrupt.

The amount $NPV(r)$ is called the *Net Present Value* of the cash flow discounted at the rate r . Given that the LendingClub repayments are monthly, the formula becomes:

$$NPV(r) = \sum_{\substack{Month=12 \times n \\ Monthi=1}}^{} \frac{CF_i}{(1 + \frac{r}{12})^i}$$

If we now have a day 1 cash flow CF_0 , we can calculate:

$$CF_0 - \sum_{\substack{Year=n \\ Yeari=1}}^{} \frac{CF_i}{(1 + r)^i}$$

However, for any given CF_0 , there is no reason that it would equal the NPV of the future cash flow (i.e no reason why the difference would be equal to zero). But this is an equation depending on r . If we can find a value of r that zeroes this formula, it is called the internal rate of return of the cash flow:

$$CF_0 - \sum_{\substack{Year=n \\ Yeari=1}}^{} \frac{CF_i}{(1 + IRR(CF))^i} = 0$$

or for monthly cash flow:

$$CF_0 - \sum_{\substack{Month=12 \times n \\ Monthi=1}}^{} \frac{CF_i}{(1 + \frac{IRR(r)}{12})^i} = 0$$

1.4 Dataset calculation

For each loan, we calculated the IRR, credit margin and NPV. The calculations were performed in Julia due to R's slow performance on such a large dataset. For this reason, the resulting datasets are available on the GitHub repository as gzipped CSV files.

1.4.1 IRR

We used the dataset to calculate the IRR of each loan. We used the following information for the dataset: `funded_amnt`(loan amount funded), `int_rate` (all-in interest rate), `term` (tenor of the loan in months), `total_pymnt` (total cumulative amount received from the borrower), `total_rec_prncp` (amount repaid allocated to principal repayment), `total_rec_int` (amount repaid allocated to interest payment), `recoveries` (any amount recovered later from the borrower) and `total_rec_late_fee` (any late payments fees paid by the borrower).

From that information, we recreated a cash flow for each loan. The R code is presented in appendix. Unfortunately, this code takes close to a full day to run on the entire dataset of completed loans (i.e. excluding all ongoing loans). This is just impractical for anybody to run to check this report and the resulting IRR results dataset is included in the Github repository. To make things practical, the dataset was actually created using code in Julia³. It is a direct translation of the R code, with a similar syntax (therefore very easy to follow). The Julia code runs about 500 times quicker (this is not a typo), or about 3 minutes. We appreciate that this is the departure from the assignment description.

Similarly, we calculate the credit margin required by each loan noting that:

³<https://julialang.org/>

$$\text{Risk-free} + \text{Credit Margin} = IRR(\text{loan})$$

1.4.2 Credit Margins

As noted in the previous section, risk-free rates change over time. When solving for the credit margin, we use the relevant risk-free rate.

Again, this was coded in Julia. The Julia code here takes about 1h20min to run. On the assumptions that the equivalent R code would take therefore almost 30 days to run through the full dataset, we did not write any R code for this calculation. The code is in appendix, and we believe easy to follow.

1.4.3 NPV

We also calculated the NPV of each loan. Again, this was coded in Julia. The code is in appendix.

We calculate both the NPV as an absolute dollar amount and as a portion of the original

Chapter 2

Dataset

The data is sourced as a *SQLite* database that downloaded from ([Kan, 2019](#)) and imported as a `tibble` `dataframe` with the `RSQlite` package. The variables were reformatted according to their respective types.

[We also sourced US zip and FIPS codes, and macroeconomical data for possible geographical statistics. The source code for the data import and reformatting is given in appendix.]

2.1 Preamble

The LendingClub dataset, although rich, is difficult to interpret. The only explanation of what the variables mean comes from a spreadsheet attached to the dataset. The explanations are not precise and/or subject to conflicting interpretation. Despite searching the LendingClub website, no further original information was found. We collected a number of reasonable assumptions in Appendix (see ([List of Assumptions](#))[list-assumptions]).

The dataset has been used a number of times in the past by various people. One paper ([Kim and Cho, 2019](#)) mentions they used a dataset that included 110 variables, which is less than ours with 145 variables. The dataset has changed over time in ways we do not know. For example, have loans been excluded because the full 145 variables were not available?

2.2 General presentation

The original dataset is large: it includes 2260668 loan samples, each containing 145 variables (after the identification variables filled with null values). The loans were issued from 2007-06-01 to 2018-12-01.

2.2.1 Business volume

The dataset represents a total of ca.\$34bln in loan principals, which is a substantial share of the total amount stated to have been intermediated to date by LC (publicly reported to be \$50bln+). About 55%/60% of the portfolio is fully repaid. See Table [2.1](#).

Figure [2.1](#) plots the number, volume (cumulative principal amount) and average principal per loan. It shows that the business grew exponentially (in the common sense of the word) from inception until 2016. At this point, according to Wikipedia ¹:

¹source: <https://en.wikipedia.org/wiki/LendingClub> - Retrieval date 15 September 2019

Table 2.1: Number of loans per status

Loan status	Count	Proportion (%)
Charged Off	261655	11.574
Current	919695	40.682
Default	31	0.001
Does not meet the credit policy. Status:Charged Off	761	0.034
Does not meet the credit policy. Status:Fully Paid	1988	0.088
Fully Paid	1041952	46.090
In Grace Period	8952	0.396
Late (16-30 days)	3737	0.165
Late (31-120 days)	21897	0.969

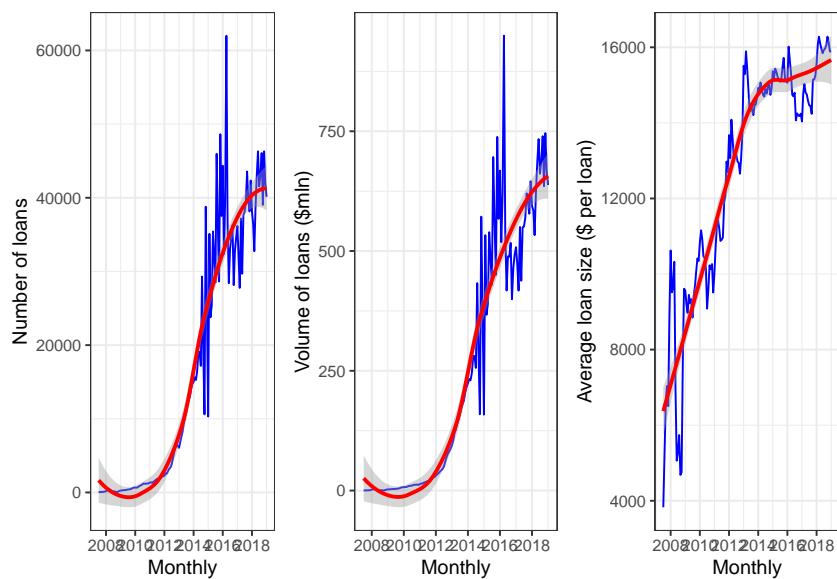


Figure 2.1: Business volume written per month

" Like other peer-to-peer lenders including Prosper, Sofi and Khutzpa.com, LendingClub experienced increasing difficulty attracting investors during early 2016. This led the firm to increase the interest rate it charges borrowers on three occasions during the first months of the year. The increase in interest rates and concerns over the impact of the slowing United States economy caused a large drop in LendingClub's share price."

The number and volume of loans plotted have been aggregated by month. The growth is very smooth in the early years, and suddenly very volatile. As far as the first part of the dataset is concerned, a starting business could expect to be volatile and could witness a yearly cycle (expected from economic consumption figures) superimposed on the growth trend. This is not the case.

An interesting metric is that the average principal of loans has increased (see RHS Figure 2.1, on a sample of 100,000 loans). Partly, the increase in the early years could be interpreted success in improving marketing, distribution capabilities and confidence building. This metric plateau-ed in 2016 and decreased afterwards, but to a much lesser extent than the gross volume metrics. However, it is more volatile than the two previous metrics in the early years.

By the end of the dataset, those metrics have essentially recovered to their 2016 level.

2.2.2 Loan lifecycle and status

In the dataset, less loans are still outstanding than matured or “*charged off*” (term that LC use to mean partially or fully written off, i.e. there are no possibilty for LC and/or the investors to receive further payments). The share of outstanding loans is:

¹ `## Share of current loans = 42.214 %`

The dataset describes the life cycle of a loan. In the typical (ideal) case, we understand it to be:

Loan is approved → Full amount funded by investors → Loan marked as Current → Fully Paid

In the worst case, it is:

Loan is approved → Full amount funded by investors → Loan marked as Current →

→ Grace period (missed payments under 2 weeks) → Late 15 to 31 days →

→ Late 31 to 120 days → Default → Charged Off

Note that *Default* precedes and is distinct from *Charged Off*². A couple of things could happen to a loan in default:

- LC and the borrower restructure the loan with a new repayment schedule, where the borrower may repay a lesser amount over a longer period; or,
- the claim could be sold to a debt recovery company that would buy the claim from LC/investors. This would be the final payment (if any) received by LC and the investors.

²See LendingClub FAQ at [and help page](#)

The dataset also describes situations where a borrower negotiated a restructuring of the repayment schedule in case of unexpected hardship (e.g. disaster, sudden unemployment).

Note that this progression of distinguishing default (event in time) from actual financial loss mirrors what banks and rating agencies do. The former is called the *Probability of Default* (PD), the latter *Loss Given Default* (LGD). Ratings change over time (in a process resembling Markov Chains). LGD show some correlations with ratings. The dataset, although detailed, does not include the full life of each loan to conduct this sort of analysis (change of loan quality over time). This is an important reason why we decided to focus on the loan approval and expected return.

CHANGE: Debt-Settlement companies ³ that explains that debt settlement companies can step in, buy the debt and pay to LC.

CHANGE: What is the difference between a loan that is in “default” and a loan that has been “charged off”? ⁴

2.2.3 Loan application

Before a loan is approved, the borrower undergoes a review process that assess his/her capacity to repay. This includes:

- employment situation and income, as well whether this income and possibly its source has been independently verified;
- whether the application is made jointly (likely with a partner or a spouse, but there are no details);
- housing situation (owner, owner with current mortgage, rental) and in which county he/she lives (that piece of information is partially anonymised by removing the last 2 digits of the borrower’s zipcode);
- the amount sought, its tenor and the purpose of the loan; and,
- what seems to be previous credit history (number of previous delinquencies). The dataset is very confusing in that regard: it is clear that such information relates to before the loan is approved in the case of the joint applicant. In the case of the principal borrower however, the variable descriptions could be read as pre-approval information, or information gathered during the life of the loan. We have assumed that the information related to the principal borrower is also pre-approval. We also used *Sales Supplements* from the LC website⁵ that describe some of the information provided to investors. LendingClub also provides a summary description of its approval process in its regulatory filings with the Securities Exchange Commission ([California, 2019](#)).

2.3 Rates

2.3.1 IRR and required credit margins

Figure 2.2 shows the evolution of credit margins over time grouped by ratings. The plots are made with a random sample of 300,000 loans.

We notice long periods where certain margins remain very stable which indicate that *both* the initial pricing was constant *and* that the proportion of default remains very low.

The graphs show considerations that are relevant to the modeling":

³<https://help.lendingclub.com/hc/en-us/articles/115011819087-Debt-settlement-companies>

⁴<https://help.lendingclub.com/hc/en-us/articles/216127747>

⁵See <https://www.lendingclub.com/legal/prospectus>

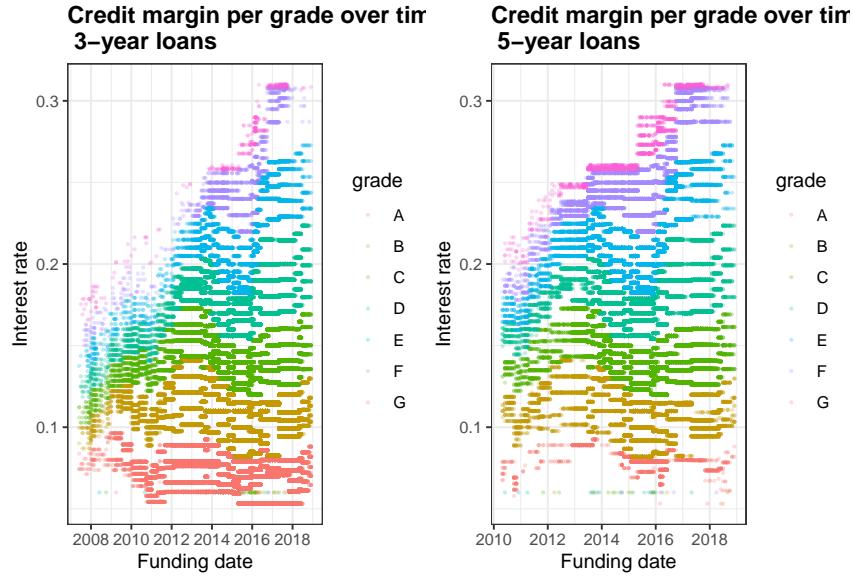


Figure 2.2: Credit margins per grade over time

- The margins clearly change over time. Predictions will require to account for time [probably in a non-linear fashion].
- For a given rating, it widens and narrows over time. The changes happen in multiples that depends on the ratings:
 - For high quality / low margin loans: the changes are multiples of the margin, for example going from roughly 3% to 6/7%.
 - Although the range of change is wide, those changes do not happen very often, especially in the later years.
 - By comparison, for low quality / high margin loans, the range of change is smaller, but more frequent and volatile.
- In other words, the relation between loan quality (its rating) and its pricing (the credit margin) will significantly non-linear.

2.3.2 Dataset

Because we are interested decisions made prior to invest, we will limit the predictors to be used to those that are realistically available prior to funding. We eliminated

2.3.3 Interest rates

Based on this information, the loan is approved or not. Approval includes the final amount (which could be lower than the amount requested), tenor (3 or 5 years) and a rating similar to those given to corporate borrowers. Unlike corporate borrowers however, the rating mechanically determines the rate of interest according to a grid known to the borrower in advance⁶. The rates have changed over time. Those changes were not as frequent as market conditions (e.g. changes in Federal Reserve Bank's rates)⁷.

Figure ??⁸ shows the predetermined interest rate depending on the initial rating as of July 2019.

⁶<https://www.lendingclub.com/investing/investor-education/interest-rates-and-fees>

⁷Corporate borrowers would negotiate interest margins on a case-by-case basis despite similar risk profiles.

⁸source: <https://www.lendingclub.com/investing/investor-education/interest-rates-and-fees>

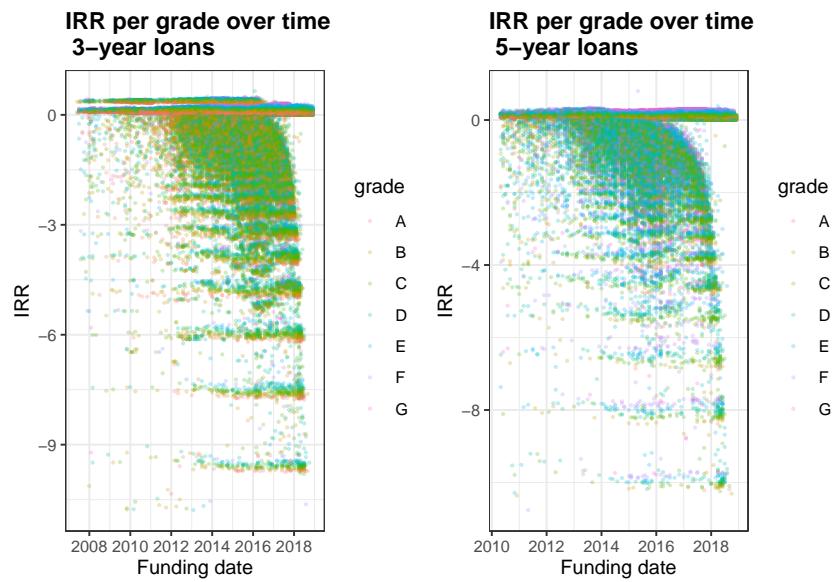


Figure 2.3: Credit margins per grade over time

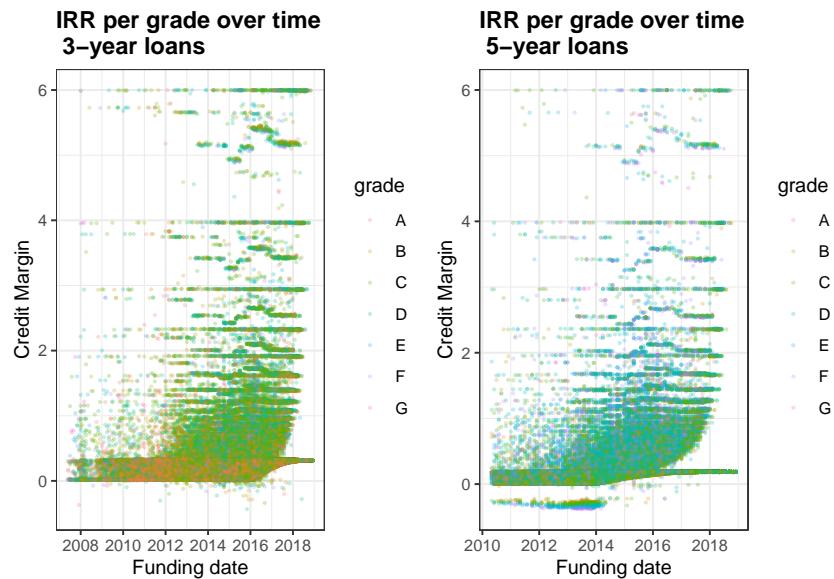


Figure 2.4: Credit margins per grade over time

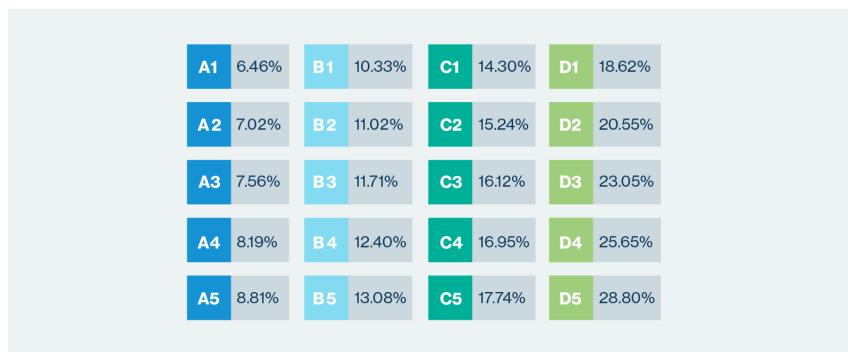


Figure 2.5: Interest rates given rating

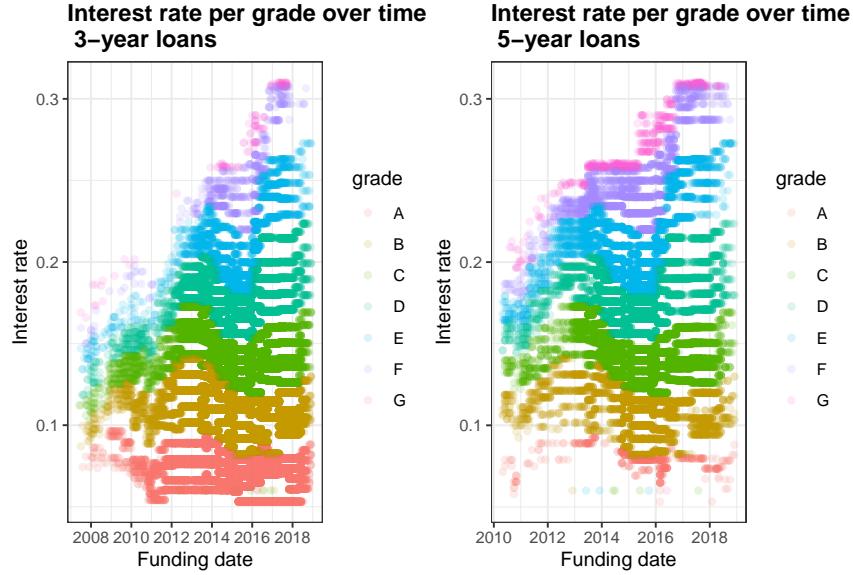


Figure 2.6: Interest rate per grade over time

At the date of this report, the ratings range from A (the best) down to D, each split in 5 sub-ratings. However, LC previously also intermediated loans rated F or G (until 6 November 2017) and E (until 30 June 2019)⁹. This explains that such ratings are in the dataset. We will assume that the ratings in the dataset are the rating at the time of approval and that, even if loans are re-rated by LC, the dataset does not reflect it.

Figures 2.6 shows the change in interest rate over time for different ratings and separated for each tenor. (Each figure is on a sample of 100,000 loans.) For each rating, we can see several parallel lines which correspond to the 5 sub-rating of each rating. We note that the range of interest rates has substantially widened over time. That is, the risk premium necessary to attract potential investors has had to substantially increase. In the most recent years, the highest rates exceed 30% which is higher than many credit cards. 3-year loans are naturally considered safer (more A-rated, less G-rated). Identical ratings attract identical rates of interest.

By comparison, we plot the 3-year (in red) and 5-year (in blue) bank swap rates in Figure 2.7. We see that the swap curve has flattened in recent times (3-year and 5-y rates are almost identical). We also can see that in broad terms the interest rates charged reflect those underlying swap rates. It is therefore most relevant to examine the credit margins added to the swap rates.

Figures 2.8 shows the change in credit margin over time for different ratings and separated for each tenor. (Each figure is on a sample of 100,000 loans.) As above, for each rating, we can see several parallel lines which correspond to the 5 sub-rating of each rating. We note that the range of credit margins has widened over time but less than the interest rates. Identical ratings attract identical credit margins.

2.3.4 Purpose

When applying, a potential borrower must state the purpose of the loan. As shown in table ??, the main purpose is the consolidation of existing debts.

⁹See <https://www.lendingclub.com/info/demand-and-credit-profile.action>

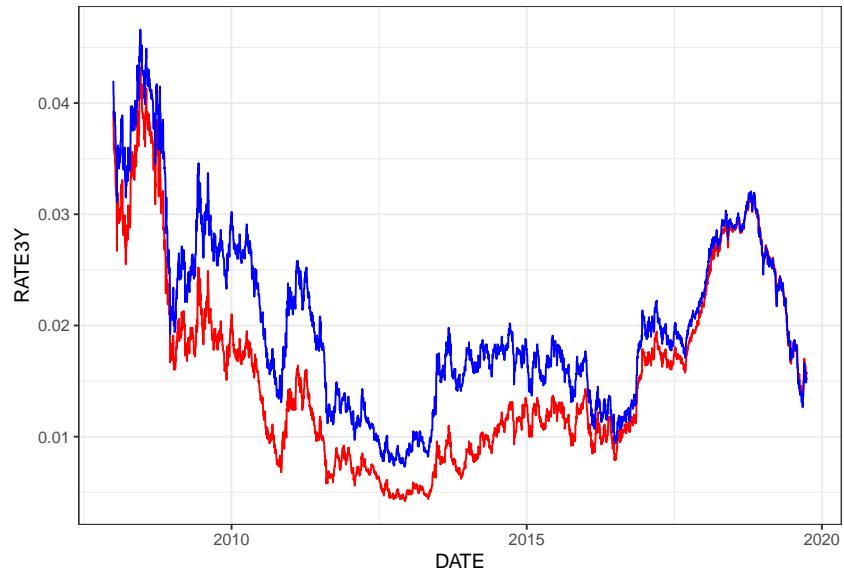


Figure 2.7: Historical Swap Rates

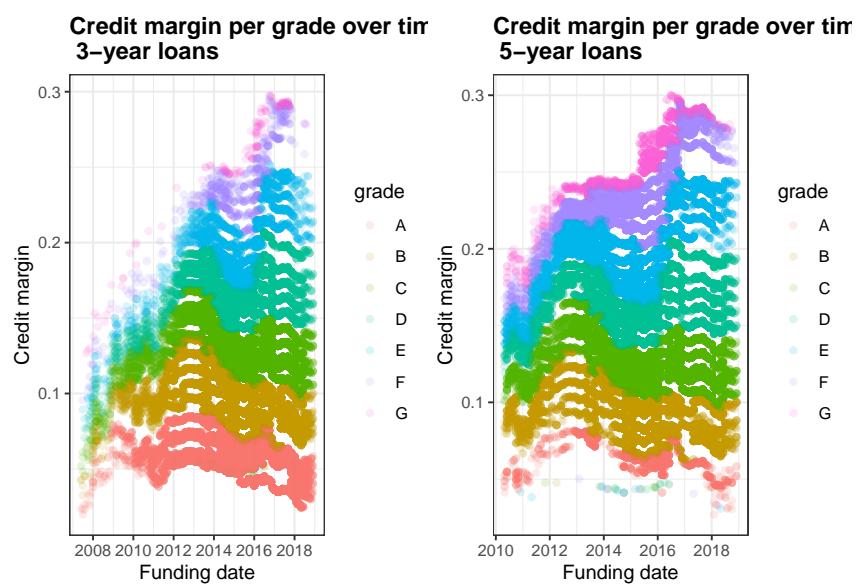


Figure 2.8: Credit margins per grade over time

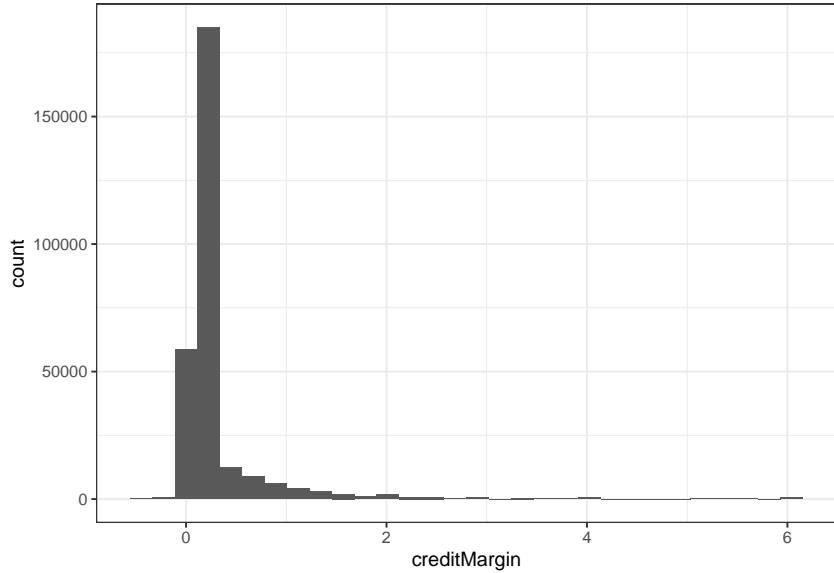


Figure 2.9: Histograms of credit margins per purpose

purpose	Number
debt_consolidation	758691
credit_card	286044
home_improvement	84709
other	75358
major_purchase	28451
small_business	15171
medical	15081
car	14184
moving	9218
vacation	8751
house	7011
wedding	2350
renewable_energy	914
educational	423

[TODO: DTI, amount... by grade]

2.3.5 Payments

The loans are approved for only two tenors, 3 and 5 years, with monthly repayments. Installments are calculated easily with the standard formula:

$$Installment = Principal \times rate \times \frac{1}{1 - \frac{1}{(1+rate)^N}}$$

Where *Principal* is the amount borrowed, *rate* = $\frac{\text{Quoted Interest Rate}}{12}$ is the monthly interest rate, and *N* is the number of installments (36 or 60 monthly payments). The following piece of code shows that the average error between this formula and the dataset value is about 2 cents. We therefore precisely understand this variable.

```

1 local({
2   installmentError <- loans %>%
3     mutate(
4       PMT = round(funded_amnt * int_rate / 12 / (1 - 1 / (1 + int_rate / 12) ^
```

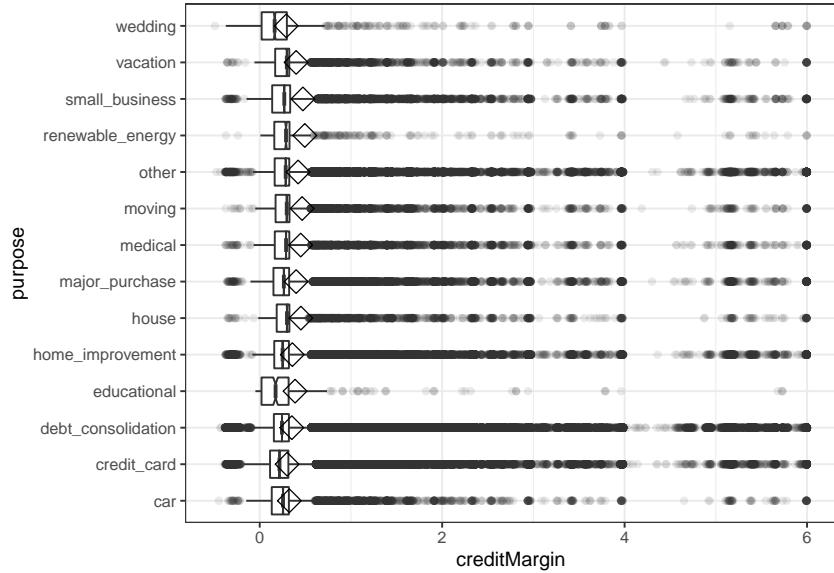


Figure 2.10: Boxplots of credit margins per purpose

```

5           term), 2),
6   PMT_delta = abs(installment - PMT)
7 ) %>%
8   select(PMT_delta)
9
10 round(mean(100 * installmentError$PMT_delta), digits = 2)
11 }

```

2.4 Net present value

The behaviour of the NPV of loan losses is important.

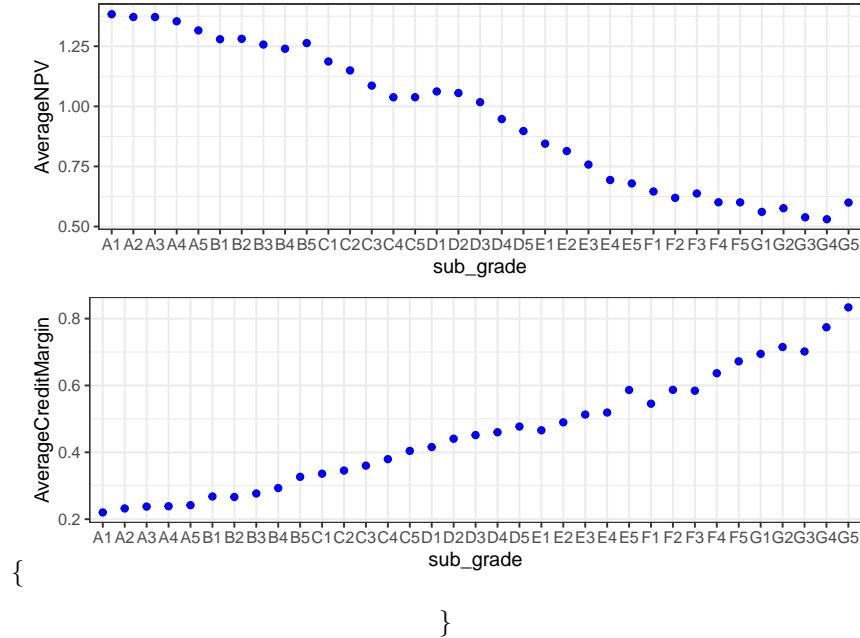
2.4.1 Average NPV and credit margin by subgrade

Figure ?? shows that as ratings worsen, the average NPV¹⁰ expressed as a portion of the funded amount decreases. For the best quality loans, we see that the NPV exceeds 1.00 = 100%: at a risk-free rate, investors receive more than what is necessary to compensate for credit loss and can use the excess to cover additional costs mentioned in the Preamble. As ratings worsen, the NPV drops down to about 50%.

If loans were adequately priced, the excess returns (thanks to higher interest) should on average offset credit losses, that is an NPV average should be at least 100%. This seems to be the case down to ratings of about D4. Further down, credit losses become too frequent and/or too substantial to be covered on average. We posit that this justified rejecting loans applications rated E1 and below.

\begin{figure}

¹⁰The averages are *not* weighted by loan amount since an investor can invest in \$25 parcels. Weighting would have been appropriate if investors were instead forced to invest in the whole amount.



\caption{Average NPV (\%) depending on sub-rating} \end{figure}

2.4.2 NPV distribution by rating

Figure 2.11 shows that for a given grade, the NPVs are very widely spread. From top to bottom, loans are grouped by ratings: from quality ratings of A and B, average ratings of C and D, to poor ratings of E and below. From left to right, we focus on different parts of how NPVs are distributed. Note that each graph is based on a random sample of 100,000 loans (about 1/12th of the original set) and therefore the NPV densities are comparable from graph to graph.

At the outset, column by column (where NPVs are on the same scale), the NPV distribution show several modes on the same location:

- a mode with a maximum around 1.25 / 1.5 being loans seemingly repaid in full (the mode is above 100% given the repayment of principal *and* interest);
- a second and third mode around 42% and -1%;
- one last very diffuse mode around -100%.

The modes are made more apparent by zooming on where the modes are present: the leftmost column basically shows the entire range of the NPVs (as portion of the loan). The middle graph zooms on the -20% / 50% range. The rightmost column zooms on the -100% / -25% section.

Looking at the left hand scale, we can see that the lower NPVs overall gain in importance as the loan rating worsen.

The overall trend is what we should expect. What is very surprising is the existence of (1) very clearly defined modes which (2) are common to all types of borrowers.

2.5 Variables

We here present other aspects of the dataset. The full list of variable is given in appendix (see Table 5.1). This dataset will be reduced as we focused on our core question: *Are LC's loans priced appropriately?*

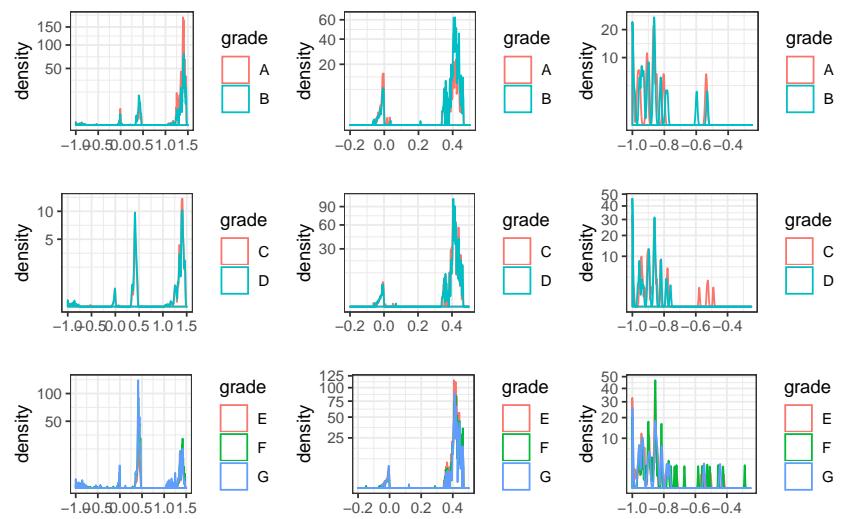


Figure 2.11: Distribution of NPV (%) depending on rating

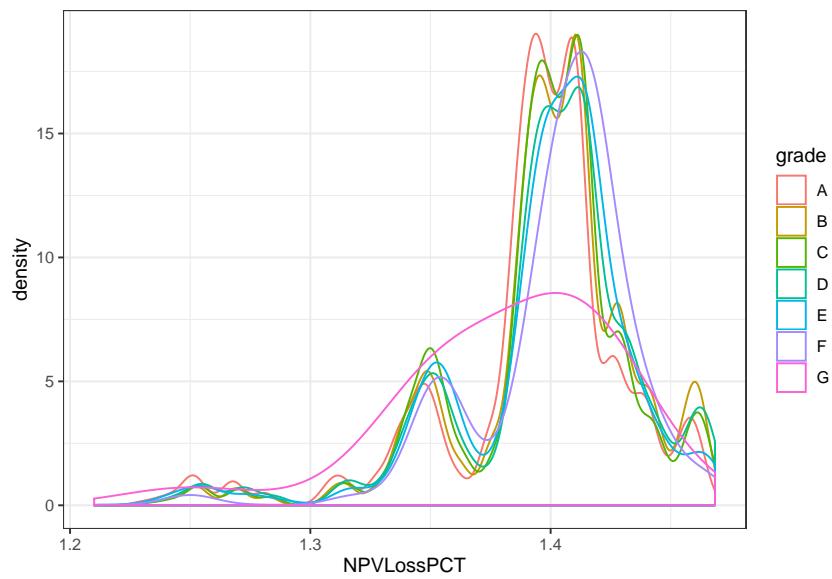


Figure 2.12: NPV % over 120% (no y-axis scaling)

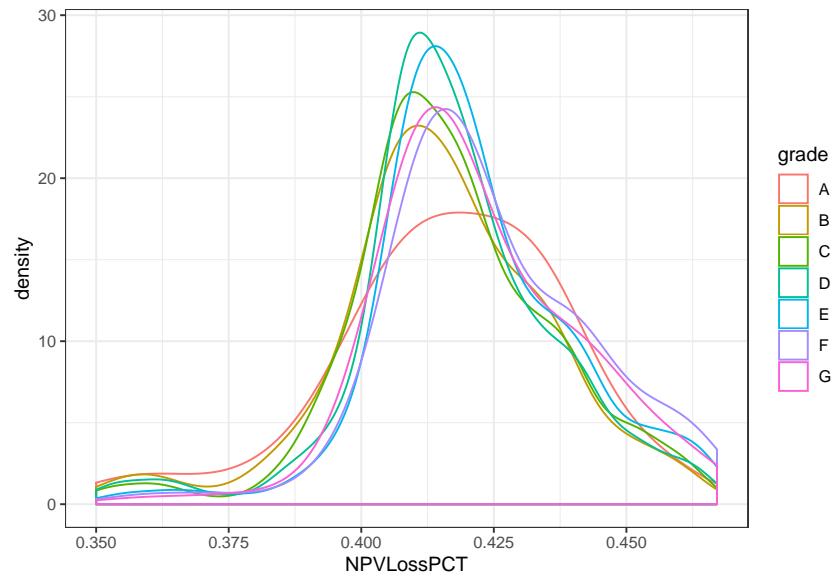


Figure 2.13: NPV % around 41% (no y-axis scaling)

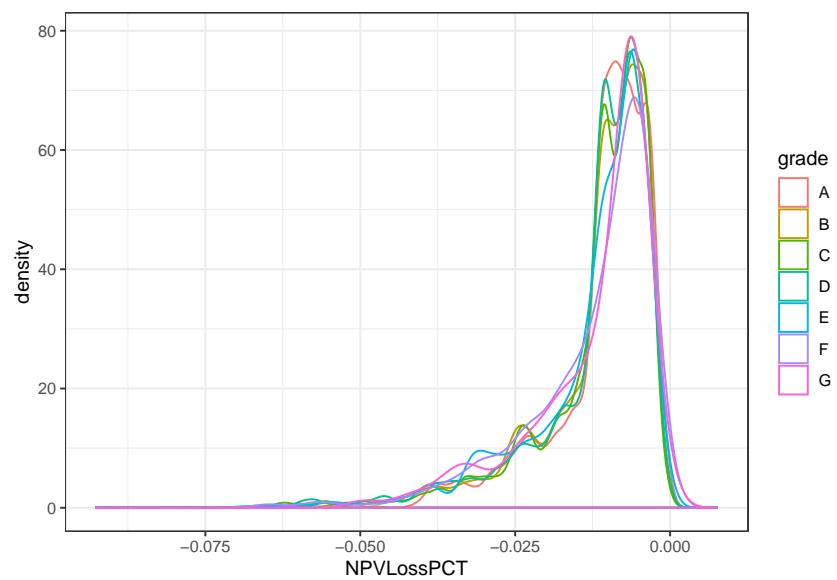


Figure 2.14: NPV % around -1% (no y-axis scaling)

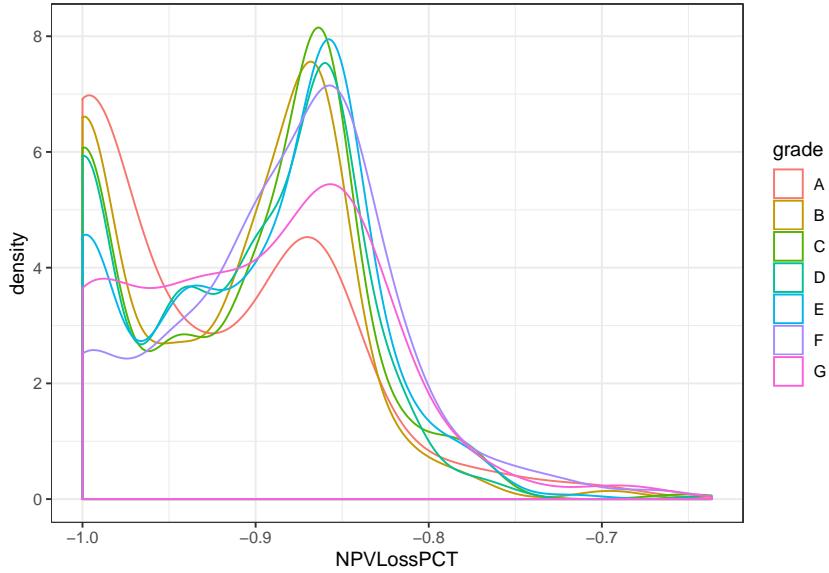


Figure 2.15: NPV % for close to total loss % (no y-axis scaling)

Table 2.2: Matured loans per status

Loan status	Count	Proportion (%)
Fully Paid	964057	24101425
Charged Off	207546	5188650
Does not meet the credit policy.	1334	33350
Status:Fully Paid		
Does not meet the credit policy.	438	10950
Status:Charged Off		

2.5.1 Identification

The dataset is anonymised (all identifying ID numbers are deleted) and we therefore removed those columns from the dataset. Since the identification IDs have been removed to anonymise the dataset, we cannot see if a borrower borrowed several times.

2.6 Loan decision

As indicated in the introduction, our focus is on loans that have gone through their entire life cycle to consider their respective pricing, risk and profitability. To that effect, we will remove all loans which are still current (either performing or not), and we will only retain loans which currently available (rated A1 to D5). Another reason for limiting ourselves to this subset is that the previous subsection graphing NPVs strongly suggests that the lowest quality loans were very underpriced yielding average losses. From here on, everything will be based on this reduced dataset.

In this reduced dataset, we focus on loans that have matured or been terminated. It contains 1306356 samples. Most of the loans (ca.80%) have been repaid in full. See Table 2.2.

When grouped by grade (Figure 2.16), we see a clear correlation between grade and default: the lower the grade the higher the portion defaults (note the limited scale with a minimum at about 50%). In addition, most of the business is written in the B- or C-rating range.

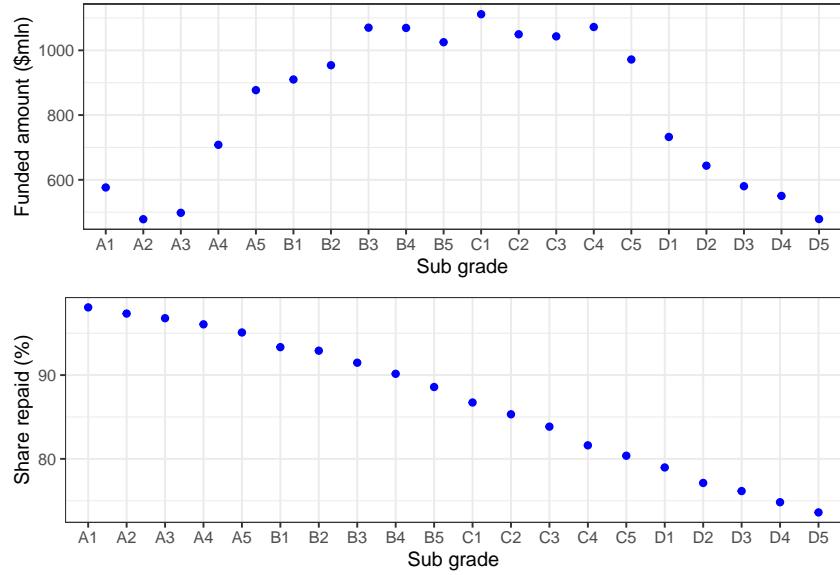


Figure 2.16: Funding and Write-offs by Sub-grades

2.7 Payment-related information

As mentioned previously, the descriptions of the dataset variables is at times incomplete or confusing. For the purpose of determining the cash flow of each individual loans, we have attempted to reconstruct the variables.

We have verified that:

- installments are calculated as per the formula shown above, rounded to the next cent.
- $\text{total_pymnt} = \text{total_rec_prncp} + \text{total_rec_int} + \text{total_rec_late_fee} + \text{recoveries}$

Chapter 3

Modelling

At the outset, the dataset presents a number of challenges:

- There is a mix of continuous and categorical data.
- The number of observations is very large.
- The possible number of predictors is large (partially due to one-hot encoding of categorical values).
- As shown in the previous section, credit margins have changed over time. This is clearly related to the wider US economic environment. Financial hardship is a key driver for some of the loans. Availability of disposable income is important to assess the ability to repay. Therefore, the cost of living, which varies from state to state, seems relevant.

[TODO: Bias/Complexity trade-off] [TODO: AoC? or AuC?]

3.1 Introduction

3.1.1 Split

The dataset was randomly split into training and validation sets (80/20 ratio).

We initiated the exploration of potential models with Principal Component Analysis, Linear Regression, Extreme Boosting and Random Forest. No model could be trained on the full set. We therefore came to limit the training set on a random sample of 0.1% (1 thousandth) of the initial full set.

Any model will require training in batch (e.g. stochastic gradient descent) or online. Our untested intuition is that online methods are not adapted to an unbalanced dataset: the number of defaults/write-offs is fairly low for high quality ratings. However, the dataset is evidently a time series which points to online training. We will not explore that line of investigation, although exploring that tension would be an interesting subject.

3.1.2 The modeling task

Two step process: - Given dataset (+/- rating?) find a score between -5 and +5 - Given a score between -5 and +5 given the right parameters for the modes - Given a density curve extrapolate the required margin

- Make money or not?

Chapter 4

Stochastic Gradient Descent

The diagram 4.1¹ shows a useful decision tree.

Following the disappointing results of the previous section, we will explore a simpler model, but iteratively trained on a wider set of random samples.

4.1 Description of the model

Input - Layer - Single Q - Distribution

4.2 Gradient descent

4.2.1 Generalities

Gradient descent is a generic numerical optimisation algorithm to iteratively converge towards a (sometimes local) minimum of a given function. It is extensively used in statistical learning to minimise error functions.

In the case of a simple linear regression model, the model training error J (the **cost function**) as a function of θ is:

$$J_{(\theta)} = \frac{1}{2} \sum_{i=1}^N \epsilon(y_i, \theta X_i)$$

where the model parameters are denoted θ_i , $X_i \in \mathbb{R}^k$ are the predictors, $Y_i \in \mathbb{R}^k$ are the responses and ϵ is a distance function. Typically, ϵ will be the Manhattan error or the Euclidian norm ($A = (a_1, \dots, a_n), B = (b_1, \dots, b_n)$).

Manhattan: $\epsilon(A, B) = \sum_{i=1}^n |a_i - b_i|$

Euclidian norm: $\epsilon(A, B) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2}$

The gradient descent algorithm uses the gradient of the error function, $\nabla J_{(\theta)}$, defined as:

$$\nabla J_{(\theta)} = \left(\frac{\partial J}{\partial \theta_0}, \frac{\partial J}{\partial \theta_1}, \dots, \frac{\partial J}{\partial \theta_p} \right)$$

¹Source: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

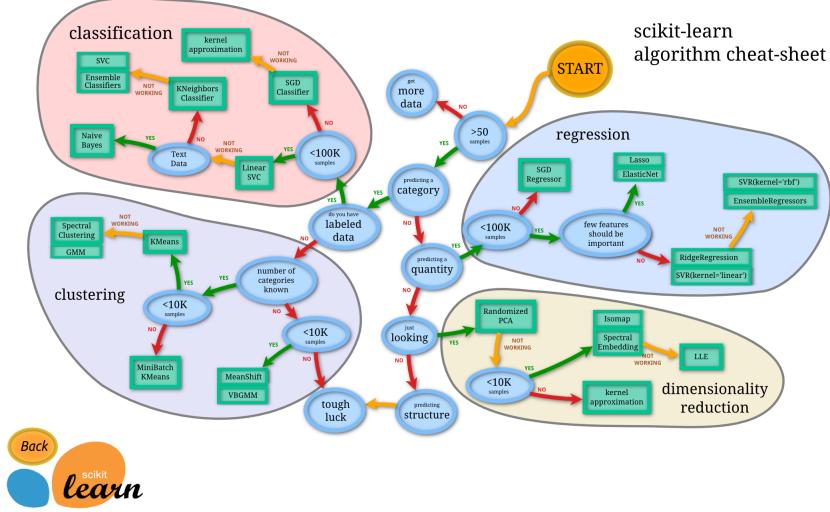


Figure 4.1: Scikit Learn algorithm cheat-sheet

And in the case of linear regression is in a matrix form that can be computed efficiently:

$$\nabla J_{(\theta)} = \left(y^T - \theta X^T \right) X$$

The gradient decent algorithm finds parameters in the following manner iterating over the training samples:

$$\text{While } ||\alpha \nabla J_{(\theta)}|| > \eta, \theta := \theta - \alpha \nabla J_{(\theta)}$$

In practice, the cost function will add a penalty term to regularise the model parameters (see below).

4.2.2 Cost function as a function of the Q parameter

Looking back at the distribution of the NPV between the -1 and about 1.5, it is multimodal and looks like the sum of 4 log-normal distributions with modes centered on about

$$PDF(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log(x)-\mu}{\sigma}\right)^2}$$

$$\text{mode} = m = e^{\mu-\sigma^2}, \text{ therefore: } \mu = \log(\text{mode}) + \sigma^2$$

$$PDF(x) = \sqrt{\frac{e}{2\pi}} \frac{1}{x\sigma} e^{-\frac{1}{2}\left(\frac{\log(\frac{x}{m})-\sigma^2}{\sigma}\right)^2}$$

We will center the distribution on the mode, therefore:

$$PDF(x) = \sqrt{\frac{e}{2\pi}} \frac{1}{(x-m)\sigma} e^{-\frac{1}{2}\left(\frac{\log(\frac{x-m}{m})-\sigma^2}{\sigma}\right)^2}$$

The distribution's tail is towards positive infinity. For the symmetric result, we would replace $(x - m)$ by $-(x - m)$.

If we use 4 log-normal distributions, the cost function is:

$$J(x, Q) = -[x - (\alpha_1 \text{PDF}_1(x, Q) + \alpha_2 \text{PDF}_2(x, Q) + \alpha_3 \text{PDF}_3(x, Q) + \alpha_4 \text{PDF}_4(x, Q))]^2$$

To optimise the shape of the total multi-modal distribution, we will assume that each α , m and σ is a linear function of Q . The derivative $\frac{\partial J}{\partial Q}$ is²:

$$\frac{\partial J}{\partial Q}(x, Q) = -\sqrt{\frac{2}{\pi}} \left(x - \frac{\alpha_1}{\sigma_1(-x + m_1)} e^{-\frac{1}{2} \left(\frac{\log \left(\frac{-x+m_1}{m_1} \right) + \sigma_1^2}{\sigma_1} \right)^2} - \frac{\alpha_2}{\sigma_2(-x + m_2)} e^{-\frac{1}{2} \left(\frac{\log \left(\frac{-x+m_2}{m_2} \right) + \sigma_2^2}{\sigma_2} \right)^2} - \frac{\alpha_3}{\sigma_3(-x + m_3)} e^{-\frac{1}{2} \left(\frac{\log \left(\frac{-x+m_3}{m_3} \right) + \sigma_3^2}{\sigma_3} \right)^2} - \frac{\alpha_4}{\sigma_4(-x + m_4)} e^{-\frac{1}{2} \left(\frac{\log \left(\frac{-x+m_4}{m_4} \right) + \sigma_4^2}{\sigma_4} \right)^2} \right)$$

4.3 Stochastic Gradient Descent

With realistic datasets, gradient descent can experience slow convergence because (1) each iteration requires calculation of the gradient for every single training example, and (2) since each individual sample is potentially very different from another, the calculated gradient may not be optimal. In such case, the gradient descent can be done using a batch of several training samples and use the average of the cost function (**batch gradient descent**). This addresses those two sources of inefficiency.

This method however still requires iterating over the entire dataset. We can instead iterate over batched of random training samples drawn from the entire dataset, instead of being drawn sequentially. This is the **stochastic gradient descent**.

Aside from the choice of the initial choice of samples, and the averaging of the cost function, the update of θ remains identical.

4.3.0.1 Cost function regularisation and derivative

Regularisation is a way to decrease the complexity of models by narrowing the range that parameters can take. It has long been established that it assists the stability and robustness of learning algorithms. See Chapter 13 of ([Shalev-Shwartz and Ben-David, 2014](#)).

Our regularised cost function is written:

[TODO]

$$J_{P,Q} = \sum_{(i,j) \in \Omega} \left(r_{i,j} - \sum_{k=1}^{N_{features}} p_{i,k} q_{j,k} \right)^2 + \frac{\lambda}{2} \left(\sum_{i,k} p_{i,k}^2 + \sum_{j,k} q_{j,k}^2 \right)$$

The gradient descent algorithm seeks to minimise the $J_{(\theta)}$ cost function by step-wise update of each model parameter x as follows:

$$\theta_{t+1}^i \leftarrow \theta_t^i - \alpha \frac{\partial J_{(\theta)}}{\partial \theta_i}$$

The parameters are the matrix coefficients $p_{i,k}$ $q_{j,k}$. α is the learning parameter that needs to be adjusted.

²This was actually generated using Maxima (code in appendix) which allows for quicker iterations.

4.3.0.2 Cost function partial derivatives

The partial derivatives of the cost function is:

$$\frac{\partial J_{P,Q}}{\partial x} = \frac{\partial}{\partial x} \left(\sum_{(i,j) \in \Omega} \left(r_{i,j} - \sum_{k=1}^{N_{features}} p_{i,k} q_{j,k} \right)^2 + \frac{\lambda}{2} \left(\sum_{i,k} p_{i,k}^2 + \sum_{j,k} q_{j,k}^2 \right) \right)$$

$$\frac{\partial J_{P,Q}}{\partial x} = \sum_{(i,j) \in \Omega} 2 \frac{\partial r_{i,j} - \sum_{k=1}^{N_{features}} p_{i,k} q_{j,k}}{\partial x} \left(r_{i,j} - \sum_{k=1}^{N_{features}} p_{i,k} q_{j,k} \right) + \frac{\lambda}{2} \left(\sum_{i,k} 2 \frac{\partial p_{i,k}}{\partial x} p_{i,k} + \sum_{j,k} 2 \frac{\partial p_{i,k}}{\partial x} q_{j,k} \right)$$

We note that $r_{i,j}$ are constants

$$\frac{\partial J_{P,Q}}{\partial x} = 2 \sum_{(i,j) \in \Omega} \sum_{k=1}^{N_{features}} \left(\frac{\partial - p_{i,k} q_{j,k}}{\partial x} \left(r_{i,j} - \sum_{k=1}^{N_{features}} p_{i,k} q_{j,k} \right) \right) + \lambda \left(\sum_{i,k} \frac{\partial p_{i,k}}{\partial x} p_{i,k} + \sum_{j,k} \frac{\partial p_{i,k}}{\partial x} q_{j,k} \right)$$

If x is a coefficient of P (resp. Q), say $p_{a,b}$ (resp. $q_{a,b}$), all partial derivatives will be nil unless for $(i,j) = (a,b)$.

Therefore:

$$\frac{\partial J_{P,Q}}{\partial p_{a,b}} = -2 \sum_{(i,j) \in \Omega} q_{j,b} \left(r_{i,j} - \sum_{k=1}^{N_{features}} p_{i,k} q_{j,k} \right) + \lambda p_{a,b}$$

and,

$$\frac{\partial J_{P,Q}}{\partial q_{a,b}} = -2 \sum_{(i,j) \in \Omega} p_{i,b} \left(r_{i,j} - \sum_{k=1}^{N_{features}} p_{i,k} q_{j,k} \right) + \lambda q_{a,b}$$

Since $\epsilon_{i,j} = r_{i,j} - \sum_{k=1}^{N_{features}} p_{i,k} q_{j,k}$ is the rating prediction error, this becomes:

$$\frac{\partial J_{P,Q}}{\partial p_{a,b}} = -2 \sum_{(i,j) \in \Omega} q_{j,b} \epsilon_{i,j} + \lambda p_{a,b}$$

and,

$$\frac{\partial J_{P,Q}}{\partial q_{a,b}} = -2 \sum_{(i,j) \in \Omega} p_{i,b} \epsilon_{i,j} + \lambda q_{a,b}$$

We note that the recent implementations of autodifferentiation algorithm (key to deep learning implementations) would avoid making those calculations by hand. We did not attempt to use the R `madness` package ³.

³<https://github.com/shabbychef/madness>

4.3.1 Stochastic Gradient Descent (SGD)

The size of the datasets is prohibitive to do those calculations across the entire training set.

Instead, we will repeatedly update the model parameters on small random samples of the training set.

Chapter 14 of (Shalev-Shwartz and Ben-David, 2014) gives an extensive introduction to various SGD algorithms.

We implemented a simple version of the algorithm and present the code in more detail.

```
1 # The algorithm is implemented from scratch and relies on nothing but the `Tidyverse` libraries.
2 library(tidyverse)
3
4 # The quality of the training and predictions is measured by the _root mean squared error_
5 # (RMSE), for which we define a few helper functions (the global variables are defined
6 # later):
7 rmse_training <- function() {
8     prediction_Z <- rowSums(LF_Model$P[tri_train$userN, ] *
9                               LF_Model$Q[tri_train$movieN, ])
10    prediction <- prediction_Z * r_sd + r_m
11    sqrt(sum((tri_train$rating - prediction) ^ 2 / nSamples))
12 }
13
14 rmse_validation <- function() {
15     prediction_Z <- rowSums(LF_Model$P[tri_test$userN, ] *
16                               LF_Model$Q[tri_test$movieN, ])
17     prediction <- prediction_Z * r_sd + r_m
18     sqrt(sum((tri_test$rating - prediction) ^ 2) / nTest)
19 }
20
21 sum_square <- function(v) {
22     return (sqrt(sum(v ^ 2) / nrow(v)))
23 }
24
25 # The key function updates the model coefficients. Its inputs are:
26 #
27 # + a list that contains the $P$ and $Q$ matrices, the training RMSE of those matrices, and
28 # a logical value indicating whether this RMSE is worse than what it was before the update
29 # (i.e. did the update diverge).
30 #
31 # + a `batch_size` that defines the number of samples to be drawn from the training set. A
32 # normal gradient descent would use the full training set; by default we only use 10,000
33 # samples out of 10 million (one tenth of a percent).
34 #
35 # + The cost regularisation `lambda` and gradient descent learning parameter `alpha`.
36 #
37 # + A number of `times` to run the descent before recalculating the RMSE and exiting the
38 # function (calculating the RMSE is computationally expensive).
39 #
40 #
41 # The training set used is less rich than the original set. As discussed, it only uses the
42 # rating (more exactly on the z_score of the rating). Genres, timestamps,... are
43 # discarded.
44
45
46 # Iterate gradient descent
47 stochastic_grad_descent <- function(model,
48                                     times = 1,
```

```

49           batch_size = 10000,
50           lambda = 0.1,
51           alpha = 0.01,
52           verbose = TRUE) {
53
54   # Run the descent `times` times.
55   for (i in 1:times) {
56
57     # Extract a sample of size `batch_size` from the training set.
58     spl <- sample(1:nSamples, size = batch_size, replace = FALSE)
59     spl_training_values <- tri_train[spl, ]
60
61     # Take a subset of `P` and `Q` matching the users and
62     # movies in the training sample.
63     spl_P <- model$P[spl_training_values$userN, ]
64     spl_Q <- model$Q[spl_training_values$movieN, ]
65
66     # rowSums returns the cross-product for a given user and movie.
67     # err is the term inside brackets in the partial derivatives
68     # calculation above.
69     err <- spl_training_values$rating_z - rowSums(spl_P * spl_Q)
70
71     # Partial derivatives wrt p and q
72     delta_P <- -err * spl_Q + lambda * spl_P
73     delta_Q <- -err * spl_P + lambda * spl_Q
74
75     model$P[spl_training_values$userN, ] <- spl_P - alpha * delta_P
76     model$Q[spl_training_values$movieN, ] <- spl_Q - alpha * delta_Q
77
78   }
79
80   # RMSE against the training set
81   error <- sqrt(sum((
82     tri_train$rating_z - rowSums(model$P[tri_train$userN, ] *
83                               model$Q[tri_train$movieN, ])
84   ) ^ 2)
85   / nSamples)
86
87   # Compares to RMSE before update
88   model$WORSE_RMSE <- (model$RMSE < error)
89   model$RMSE <- error
90
91   # Print some information to keep track of success
92   if (verbose) {
93     cat(
94       "# features=",
95       ncol(model$P),
96       " J=",
97       nSamples * error ^ 2 +
98         lambda / 2 * (sum(model$P ^ 2) + sum(model$Q ^ 2)),
99       " Z-scores RMSE=",
100      model$RMSE,
101      "\n"
102    )
103    flush.console()
104  }
105
106  return(model)

```

```

107
108
109 rm(list_results)
110 list_results <- tibble(
111   "alpha" = numeric(),
112   "lambda" = numeric(),
113   "nFeatures" = numeric(),
114   "rmse_training_z_score" = numeric(),
115   "rmse_training" = numeric(),
116   "rmse_validation" = numeric()
117 )
118
119 # The main training loop runs as follows:
120 #
121 # + We start with 3 features.
122 #
123 # + The model is updated in batches of 100 updates. This is done up to 250 times. At each
124 # time, if the model starts diverging, the learning parameter ( $\alpha$ ) is reduced.
125 #
126 # + Once the 250 times have passed, or if  $\alpha$  has become incredibly small, or if the
127 # RMSE doesn't really improve anymore (by less than 1 millionth), we add another features
128 # and start again.
129
130 initial_alpha <- 0.1
131 for (n in 1:100) {
132   # Current number of features
133   number_features <- ncol(LF_Model$P)
134
135   # lambda = 0.01 for 25 features, i.e. for about 2,000,000 parameters.
136   # We keep lambda proportional to the number of features
137   lambda <- 0.1 * (nUsers + nMovies) * number_features / 2000000
138
139   alpha <- initial_alpha
140
141   cat(
142     "CURRENT FEATURES: ",
143     number_features,
144     "---- Pre-training validation RMSE = ",
145     rmse_validation(),
146     "\n"
147   )
148
149   list_results <- list_results %>% add_row(
150     alpha = alpha,
151     lambda = lambda,
152     nFeatures = number_features,
153     rmse_training_z_score = LF_Model$RMSE,
154     rmse_training = rmse_training(),
155     rmse_validation = rmse_validation()
156   )
157
158   for (i in 1:250) {
159     pre_RMSE <- LF_Model$RMSE
160     LF_Model <- stochastic_grad_descent(
161       model = LF_Model,
162       times = 100,
163       batch_size = 1000 * number_features,
164       alpha = alpha,

```

```

165     lambda = lambda
166   )
167
168   list_results <- list_results %>% add_row(
169     alpha = alpha,
170     lambda = lambda,
171     nFeatures = number_features,
172     rmse_training_z_score = LF_Model$RMSE,
173     rmse_training = rmse_training(),
174     rmse_validation = rmse_validation()
175   )
176
177   if (LF_Model$WORSE_RMSE) {
178     alpha <- alpha / 2
179     cat("Decreasing gradient parameter to: ", alpha, "\n")
180   }
181
182   if (initial_alpha / alpha > 1000 |
183     abs((LF_Model$RMSE - pre_RMSE) / pre_RMSE) < 1e-6) {
184     break()
185   }
186 }
187
188
189 # RMSE against validation set:
190 rmse_validation_post <- rmse_validation()
191 cat(
192   "CURRENT FEATURES: ",
193   number_features,
194   "---- POST-training validation RMSE = ",
195   rmse_validation_post,
196   "\n"
197 )
198
199 # if (number_features == 12){
200 #   break()
201 # }
202
203
204 # Add k features
205 k_features <- 1
206 LF_Model$P <- cbind(LF_Model$P,
207   matrix(
208     rnorm(
209       nrow(LF_Model$P) * k_features,
210       mean = 0,
211       sd = sd(LF_Model$P) / 100
212     ),
213     nrow = nrow(LF_Model$P),
214     ncol = k_features
215   ))
216
217 LF_Model$Q <- cbind(LF_Model$Q,
218   matrix(
219     rnorm(
220       nrow(LF_Model$Q) * k_features,
221       mean = 0,
222       sd = sd(LF_Model$Q) / 100

```

```
223     ),
224     nrow = nrow(LF_Model$Q),
225     ncol = k_features
226   ))
227
228 }
229
230 saveRDS(list_results, "datasets/LRMF_results.rds")
```

Chapter 5

Conclusion

Is the interest rate proposed by LC high enough?

We use the entire dataset to estimate the impact of time as a polynomial curve. To assess future loans, this would not be acceptable. Only an online algorithm should be used. For example ARMA/ARIMA, Kalman filtering of the time-trend trajectory.

This report was one of those “it-builds-character” endeavour. Facing such a large dataset, it took a very long time before settling to address a particular question. Exploring the data lead to many blind alleys with little interest.

Appendix

5.1 List of assumptions / limitations regarding the dataset

As mentioned during this report, we had to make numerous assumptions given the lack of clarity of the variable descriptions.

- *Dataset quality:* Aside from cents rounding issues, the dataset does not contain any flagrant errors that we could see (e.g. minor error of amount or rate, zipcode). Quality of the variable description is a different matter altogether.
- *Ratings:* The day-1 rating is between A1 and (and no lower than) G5. No note is rated lower than E5 after 6 November 2017, and lower than D5 after 30 June 2019.
- *Credit history:* Credit history information for the principal borrower relates to pre-approval and not post-funding. This is clear for the joint applicants, but simply an assumption for the principal borrower.
- *Recoveries:* Recoveries (if any) are assumed to be paid 3 months after the last scheduled payment date (variable `last_pymnt_d`)
- *Survival effect:* The dataset does not include applications that were rejected by the lender (for whatever reason) or by the borrower (for example because the interest rate quote is too high). It may also be the case that some actual loans were excluded as and when the dataset changed over the years.
- LIBOR funding rate: we use the 3-year and 5-year swap rates. In reality, we should have used average tenor-weighted swap rates (i.e. ca. 1.5 Y and 2.5 Y). This requires a full swap curve and more calculation than necessary for our purpose. The principles of this report should not be significantly affected by this approximation.

We do hope that LendingClub investors receive information of much better quality!

5.2 Data preparation and formatting

We used different sources of information:

- The LendingClub dataset made available on Kaggle;
- US geographical data about zip and FIPS codes;
- Market interest rates from the Saint Louis Federal Reserve Bank; and,
- Macro data from the same source.

We here show the code used to prepare the data. It was automatically formatted by *RStudio*.

5.2.1 LendinClub dataset

```
1 local({
2   #
3   # STEP 1: Download the dataset
4   #
5   # Got to https://www.kaggle.com/wendykan/lending-club-loan-data
6   #
7   # Download into the 'datasets' subdirectory
8   # Unzip the file.
9   # WARNING: The unzipping will be about 2.4GB
10  #
11  # Name the sql database "datasets/lending_club.sqlite"
12  #
13  #
14  # STEP 2: Prepare the database as a tibble
15  #
16  #
17  ##
18  ## WARNING: THIS ASSUMES A 'datasets' DIRECTORY WAS CREATED
19  ##
20  library(RSQLite)
21  db_conn <-
22    dbConnect(RSQLite::SQLite(), "datasets/lending_club.sqlite")
23  dbListTables(db_conn)
24
25
26  # Returns a 2.96GB data frame
27  lending_club <- dbGetQuery(db_conn, "SELECT * FROM loan")
28  lending_club <- as_tibble(lending_club)
29
30  # Close the database
31  dbDisconnect(db_conn)
32
33  # Compressed to ca.285MB on disk
34  saveRDS(lending_club, "datasets/lending_club.rds")
35
36
37  library(tidyverse)
38  library(lubridate)
39  library(hablar)
40
41  # Before reformat in case the previous step was already done
42  # lending_club <- readRDS("datasets/lending_club.rds")
43  #
44  # str(lending_club)
45  #
46
47  # Leave the original dataset untouched and work with a copy.
48  lc <- lending_club
49
50  lc <- lc %>%
51
52  # Add loan identification number to track the loans across calculations
53  mutate(loanID = row_number()) %>%
54
55  # Remove useless strings
56  mutate(
```

```

57     term      = str_remove(term, " months"),
58     emp_length = str_replace(emp_length, "<1", "0"),
59     emp_length = str_replace(emp_length, "10+", "10"),
60     emp_length = str_remove(emp_length, "years")
61 ) %>%
62
63 # Creates dates out of strings - Parse errors will be raised when no dates.
64 mutate(
65   debt_settlement_flag_date = as_date(dmy(
66     str_c("1-", debt_settlement_flag_date)
67   )),
68   earliest_cr_line        = as_date(dmy(str_c(
69     "1-", earliest_cr_line
70   ))),
71   hardship_start_date     = as_date(dmy(str_c(
72     "1-", hardship_start_date
73   ))),
74   hardship_end_date       = as_date(dmy(str_c(
75     "1-", hardship_end_date
76   ))),
77   issue_d                 = as_date(dmy(str_c("1-", issue_d))),
78   last_credit_pull_d      = as_date(dmy(str_c(
79     "1-", last_credit_pull_d
80   ))),
81   last_pymnt_d            = as_date(dmy(str_c(
82     "1-", last_pymnt_d
83   ))),
84   next_pymnt_d            = as_date(dmy(str_c(
85     "1-", next_pymnt_d
86   ))),
87   payment_plan_start_date = as_date(dmy(str_c(
88     "1-", payment_plan_start_date
89   ))),
90   sec_app_earliest_cr_line = as_date(dmy(str_c(
91     "1-", sec_app_earliest_cr_line
92   ))),
93   settlement_date          = as_date(dmy(str_c(
94     "1-", settlement_date
95   ))))
96 ) %>%
97
98 # Bulk type conversion with convert from the `hablar` package
99 convert(
100   # Strings
101   chr(emp_title, title, url, zip_code),
102
103   # Factors
104   fct(
105     addr_state,
106     application_type,
107     debt_settlement_flag,
108     desc,
109     disbursement_method,
110     grade,
111     hardship_flag,
112     hardship_loan_status,
113     hardship_reason,
114     hardship_status,

```

```

115     hardship_type,
116     home_ownership,
117     id,
118     initial_list_status,
119     loan_status,
120     member_id,
121     policy_code,
122     purpose,
123     pymnt_plan,
124     settlement_status,
125     sub_grade,
126     verification_status,
127     verification_status_joint
128 ),
129
130 # Integers
131 int(
132     acc_now_delinq,
133     acc_open_past_24mths,
134     chargeoff_within_12_mths,
135     collections_12_mths_ex_med,
136     deferral_term,
137     delinq_2yrs,
138     emp_length,
139     hardship_dpd,
140     hardship_length,
141     inq_fi,
142     inq_last_12m,
143     inq_last_6mths,
144     mo_sin_old_il_acct,
145     mo_sin_old_rev_tl_op,
146     mo_sin_rcnt_rev_tl_op,
147     mo_sin_rcnt_tl,
148     mort_acc,
149     mths_since_last_delinq,
150     mths_since_last_major_derog,
151     mths_since_last_record,
152     mths_since_rcnt_il,
153     mths_since_recent_bc,
154     mths_since_recent_bc_dlq,
155     mths_since_recent_inq,
156     mths_since_recent_revol_delinq,
157     num_accts_ever_120_pd,
158     num_actv_bc_tl,
159     num_actv_rev_tl,
160     num_bc_sats,
161     num_bc_tl,
162     num_il_tl,
163     num_op_rev_tl,
164     num_rev_accts,
165     num_rev_tl_bal_gt_0,
166     num_sats,
167     num_tl_120dpd_2m,
168     num_tl_30dpd,
169     num_tl_90g_dpd_24m,
170     num_tl_op_past_12m,
171     open_acc,
172     open_acc_6m,

```

```

173     open_act_il,
174     open_il_12m,
175     open_il_24m,
176     open(rv_12m,
177     open(rv_24m,
178     sec_app_chargeoff_within_12_mths,
179     sec_app_collections_12_mths_ex_med,
180     sec_app_inq_last_6mths,
181     sec_app_mort_acc,
182     sec_app_mths_since_last_major_derog,
183     sec_app_num_rev_accts,
184     sec_app_open_acc,
185     sec_app_open_act_il,
186     term
187   ),
188
189   # Floating point
190   dbl(
191     all_util,
192     annual_inc,
193     annual_inc_joint,
194     avg_cur_bal,
195     bc_open_to_buy,
196     bc_util,
197     collection_recovery_fee,
198     delinq_amnt,
199     dti,
200     dti_joint,
201     funded_amnt,
202     funded_amnt_inv,
203     hardship_amount,
204     hardship_last_payment_amount,
205     hardship_payoff_balance_amount,
206     il_util,
207     installment,
208     int_rate,
209     last_pymnt_amnt,
210     loan_amnt,
211     max_bal_bc,
212     orig_projected_additional_accrued_interest,
213     out_prncp,
214     out_prncp_inv,
215     pct_tl_nvr_dlq,
216     percent_bc_gt_75,
217     pub_rec,
218     pub_rec_bankruptcies,
219     recoveries,
220     revol_bal,
221     revol_bal_joint,
222     revol_util,
223     sec_app_revol_util,
224     settlement_amount,
225     settlement_percentage,
226     tax_liens,
227     tot_coll_amt,
228     tot_cur_bal,
229     tot_hi_cred_lim,
230     total_acc,

```

```

231     total_bal_ex_mort,
232     total_bal_il,
233     total_bc_limit,
234     total_cu_tl,
235     total_il_high_credit_limit,
236     total_pymnt,
237     total_pymnt_inv,
238     total_rec_int,
239     total_rec_late_fee,
240     total_rec_prncp,
241     total_rev_hi_lim
242   )
243 ) %>%
244
245 # Converts some values to 1/-1 (instead of Boolean)
246 mutate(
247   pymnt_plan = if_else(pymnt_plan == "y", 1, -1),
248   hardship_flag = if_else(hardship_flag == "Y", 1, -1),
249   debt_settlement_flag = if_else(debt_settlement_flag == "Y", 1, -1)
250 ) %>%
251
252 # Some values are percentages
253 mutate(
254   int_rate = int_rate / 100,
255   dti = dti / 100,
256   dti_joint = dti_joint / 100,
257   revol_util = revol_util / 100,
258   il_util = il_util / 100,
259   all_util = all_util / 100,
260   bc_open_to_buy = bc_util / 100,
261   pct_tl_nvr_dlq = pct_tl_nvr_dlq / 100,
262   percent_bc_gt_75 = percent_bc_gt_75 / 100,
263   sec_app_revol_util = sec_app_revol_util / 100
264 ) %>%
265
266 # Create quasi-centered numerical grades out of grade factors with "A" = 3 down to "G" = -3
267 mutate(grade_num = 4 - as.integer(grade)) %>%
268
269 # Ditto with sub_grades. "A1" = +3.4, "A3" = +3.0, down to "G3" = -3.0, "G5" = -3.4
270 mutate(sub_grade_num = 3.6 - as.integer(sub_grade) / 5) %>%
271
272 # Keep the first 3 digits of the zipcode as numbers
273 mutate(zip_code = as.integer(str_sub(zip_code, 1, 3))) %>%
274
275 # order by date
276 arrange(issue_d) %>%
277
278 # Remove empty columns
279 select(-id, -member_id, -url)
280
281 saveRDS(lc, "datasets/lending_club_reformatted.rds")
282
283
284 # Select loans which have matured or been terminated
285 past_loans <- lc %>%
286   filter(
287     loan_status %in% c(
288       "Charged Off",

```

```

289     "Does not meet the credit policy. Status:Charged Off",
290     "Does not meet the credit policy. Status:Fully Paid",
291     "Fully Paid"
292   )
293 )
294
295 saveRDS(past_loans, "datasets/lending_club_reformatted_paid.rds")
296 }```

```

5.2.2 Zip codes and FIPS codes

The R package `zipcode` was installed.

```

1  #
2  # ZIPCodes dataset.
3  #
4
5  library(zipcode)
6  data(zipcode)
7  zips <- zipcode %>%
8    as_tibble() %>%
9    mutate(zip = as.integer(str_sub(zip, 1, 3)))
10
11 saveRDS(zips, "datasets/zips.rds")

```

A csv file containing zip codes, FIPS codes and population information was downloaded from the *Simple Maps*¹ website.

```

1 local({
2   kaggleCodes <- read.csv("datasets/csv/ZIP-COUNTY-FIPS_2017-06.csv")
3
4   kaggleCodes <-
5     kaggleCodes %>%
6     as_tibble() %>%
7     mutate(zip = floor(ZIP/100),
8           FIPS = STCOUNTYFP,
9           COUNTYNAME = str_replace(COUNTYNAME, pattern = "County", replacement = ""),
10          COUNTYNAME = str_replace(COUNTYNAME, pattern = "Borough", replacement = ""),
11          COUNTYNAME = str_replace(COUNTYNAME, pattern = "Municipio", replacement = ""),
12          COUNTYNAME = str_replace(COUNTYNAME, pattern = "Parish", replacement = ""),
13          COUNTYNAME = str_replace(COUNTYNAME, pattern = "Census Area", replacement = "")) %>%
14     rename(county = COUNTYNAME) %>%
15     select(zip, county, FIPS) %>%
16     arrange(zip)
17
18   saveRDS(zipfips, "datasets/kaggleCodes.rds")
19 })

```

5.2.3 Market interest rates

Market interest rates (3-year and 5-year swap rates) were download from the Saint Louis Federal Reserve Bank. Datasets are split between before and after the LIBOR fixing scandal. The datasets are merged with disctinct dates.

Download sources are:

- Pre-LIBOR 3-y swap <https://fred.stlouisfed.org/series/DSWP3>

¹<https://simplemaps.com/data/us-zips>

- Post-LIBOR 3-y swap <https://fred.stlouisfed.org/series/ICERATES1100USD3Y>
- Pre-LIBOR 5-y swap <https://fred.stlouisfed.org/series/MSWP5>
- Post-LIBOR 5-y swap <https://fred.stlouisfed.org/series/ICERATES1100USD5Y>

```

1 local({
2   LIBOR3Y <- read.csv("datasets/csv/DSWP3.csv") %>%
3     as_tibble() %>%
4     filter(DSWP3 != ".") %>%
5     mutate(DATE = as_date(DATE),
6           RATE3Y = as.numeric(as.character(DSWP3)) / 100) %>%
7     select(DATE, RATE3Y)
8
9   ICE3Y <- read.csv("datasets/csv/ICERATES1100USD3Y.csv") %>%
10    as_tibble() %>%
11    filter(ICERATES1100USD3Y != ".") %>%
12    mutate(DATE = as_date(DATE),
13           RATE3Y = as.numeric(as.character(ICERATES1100USD3Y)) / 100) %>%
14    select(DATE, RATE3Y)
15
16
17   LIBOR5Y <- read.csv("datasets/csv/DSWP5.csv") %>%
18     as_tibble() %>%
19     filter(DSWP5 != ".") %>%
20     mutate(DATE = as_date(DATE),
21           RATE5Y = as.numeric(as.character(DSWP5)) / 100) %>%
22     select(DATE, RATE5Y)
23
24   ICE5Y <- read.csv("datasets/csv/ICERATES1100USD5Y.csv") %>%
25     as_tibble() %>%
26     filter(ICERATES1100USD5Y != ".") %>%
27     mutate(DATE = as_date(DATE),
28           RATE5Y = as.numeric(as.character(ICERATES1100USD5Y)) / 100) %>%
29     select(DATE, RATE5Y)
30
31   RATES3Y <- LIBOR3Y %>% rbind(ICE3Y) %>%
32     arrange(DATE) %>% distinct(DATE, .keep_all = TRUE)
33
34   RATES5Y <- LIBOR5Y %>% rbind(ICE5Y) %>%
35     arrange(DATE) %>% distinct(DATE, .keep_all = TRUE)
36
37   saveRDS(RATES3Y, "datasets/rates3Y.rds")
38   saveRDS(RATES5Y, "datasets/rates5Y.rds")
39
40
41   # Note there are 7212 days from 1 Jan 2000 to 30 Sep 2019
42   #
43   # (ymd("2000-01-01") %--% ymd("2019-09-30")) %/%
44   # days(1)
45   RATES <- tibble(n = seq(0, 7212)) %>%
46
47   # Create a column with all dates
48   mutate(DATE = ymd("2000-01-01") + days(n)) %>%
49   select(-n) %>%
50
51   # Add all daily 3- then 5-year rates and fill missing down
52   left_join(RATES3Y) %>%
53   fill(RATE3Y, .direction = "down") %>%

```

```

54   left_join(RATES5Y) %>%
55     fill(RATE5Y, .direction = "down")
56
57   saveRDS(RATES, "datasets/rates.rds")
58 }

```

5.2.4 Macro-economical data

Macro-economical datasets were sourced from the same website as Microsoft Excel files. They were converted as-is to tab-separated csv files with LibreOffice.

- Median income per household: <https://geofred.stlouisfed.org/map/?th=pubugn&cc=5&rc=false&im=fractile&sb&lng=-112.41&lat=44.31&zsm=4&sl&sv&am=Average&at=Not%20Seasonally%20Adjusted,%20Annual,%20Dollars&sti=2022&fq=Annual&rt=county&un=lin&dt=2017-01-01>
- Per capita personal income: <https://geofred.stlouisfed.org/map/?th=pubugn&cc=5&rc=false&im=fractile&sb&lng=-112.41&lat=44.31&zsm=4&sl&sv&am=Average&at=Not%20Seasonally%20Adjusted,%20Annual,%20Dollars&sti=882&fq=Annual&rt=county&un=lin&dt=2017-01-01>
- Unemployment: <https://geofred.stlouisfed.org/map/?th=rdpu&cc=5&rc=false&im=fractile&sb&lng=-90&lat=40&zsm=4&sl&sv&am=Average&at=Not%20Seasonally%20Adjusted,%20Monthly,%20Percent&sti=1224&fq=Monthly&rt=county&un=lin&dt=2019-08-01>

```

1 local({
2 #####
3 ## Median income per household by FIPS from 2002 to 2017
4 ##
5 # Prepare median income
6 medianIncome <-
7   # Load the dataset after dropping the first line
8   read.csv(
9     "datasets/csv/GeoFRED_Estimate_of_Median_Household_Income_by_County_Dollars.csv",
10    sep = "\t",
11    skip = 1,
12    stringsAsFactors = FALSE
13  ) %>%
14
15  # Drops columns containing a unique identifier and the FIPS name
16  select(-"Series.ID", -"Region.Name") %>%
17
18  # Rename the relevant column to 'FIPS'
19  rename(FIPS = "Region.Code") %>%
20
21  # Order by FIPS
22  arrange(FIPS) %>%
23
24  # Convert to a 'long' table, i.e. one column for FIPS, one for date, one for income
25  pivot_longer(cols = starts_with("X"),
26                names_to = "Date",
27                values_to = "medianIncome") %>%
28
29  # Create actual dates
30  mutate(Date = str_replace(Date, "[X]", ""))
31

```

```

32         Date = ymd(str_c(Date, "-12-31")))
33 
34 saveRDS(medianIncome, "datasets/medianincome.rds")
35 
36 
37 #####
38 #####
39 ## 
40 ## Per capita income by FIPS from 2002 to 2017
41 ##
42 personalIncome <-
43 # Load the dataset after dropping the first line
44 read.csv(
45   "datasets/csv/GeoFRED_Per_Capita_Personal_Income_by_County_Dollars.csv",
46   sep = "\t",
47   skip = 1,
48   stringsAsFactors = FALSE
49 ) %>%
50 
51 # Drops columnsn containing a unique identifier and the FIPS name
52 select(-"Series.ID", -"Region.Name") %>%
53 
54 # Rename the relevant column to 'FIPS'
55 rename(FIPS = "Region.Code") %>%
56 
57 # Order by FIPS
58 arrange(FIPS) %>%
59 
60 # Convert to a 'long' table, i.e. one column for FIPS, one for date, one for income
61 pivot_longer(cols = starts_with("X"),
62               names_to = "Date",
63               values_to = "personalIncome") %>%
64 
65 # Create actual dates
66 mutate(Date = str_replace(Date, "[X]", ""),
67        Date = ymd(str_c(Date, "-12-31")))
68 
69 saveRDS(personalIncome, "datasets/personalincome.rds")
70 
71 #####
72 ## 
73 ## 
74 ## Unemployment rate monthly by FIPS from January 2000 to August 2019
75 ##
76 unemploymentRate <-
77 # Load the dataset after dropping the first line
78 read.csv(
79   "datasets/csv/GeoFRED_Unemployment_Rate_by_County_Percent.csv",
80   sep = "\t",
81   skip = 1,
82   stringsAsFactors = FALSE
83 ) %>%
84 
85 # Drops columnsn containing a unique identifier and the FIPS name
86 select(-"Series.ID", -"Region.Name") %>%
87 
88 # Rename the relevant column to 'FIPS'
89 rename(FIPS = "Region.Code") %>%

```

```

90
91 # Order by FIPS
92 arrange(FIPS) %>%
93
94 mutate_all(as.double) %>%
95
96 # Convert to a 'long' table, i.e. one column for FIPS, one for date, one for income
97 pivot_longer(cols = starts_with("X"),
98             names_to = "Date",
99             values_to = "unemploymentRate",
100            values_ptypes = c("unemploymentRate", numeric)) %>%
101
102 # Converts the content of the Year column to an actual date
103 mutate(
104   Date = str_replace(Date, "[X]", ""),
105   Date = str_replace(Date, "[.]", "-"),
106   Date = ymd(str_c(Date, "-1")))
107 )
108
109 saveRDS(unemploymentRate, "datasets/unemployment.rds")
110 }

```

5.3 List of variables

This table presents the list of variables provided in the original dataset. The descriptions come from a spreadsheet attached with the dataset and, unfortunately, are not extremely precise and subject to interpretation. We added comments and/or particular interpretations in *CAPITAL LETTERS*.

Table 5.1: Description of the dataset variables as provided in the dataset downloaded from Kaggle

Variable Name	Description
loanID	NOTE THIS IS NOT AN ORIGINAL VARIABLE. IT WAS ADDED FOR THE PURPOSE OF TRACKING LOANS INDIVIDUALLY AS AND WHEN NEEDED.
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.
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
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.

Table 5.1: Description of the dataset variables as provided in the dataset downloaded from Kaggle (*continued*)

Variable Name	Description
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, NONE
annual_inc	The self-reported annual income provided by the borrower during registration. NOT USED AS A VARIABLE SINCE JOINT INCOME ALREADY INCLUDES IT.
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
pymnt_plan	Indicates if a payment plan has been put in place for the loan
url	URL for the LC page with listing data.
desc	Loan description provided by the borrower
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. NOT USED AS A VARIABLE. ONLY USE JOINT DTI.
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
earliest_cr_line	The month the borrower's earliest reported credit line was opened
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
mths_since_last_delinq	The number of months since the borrower's last delinquency.
mths_since_last_record	The number of months since the last public record.
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.

Table 5.1: Description of the dataset variables as provided in the dataset downloaded from Kaggle (*continued*)

Variable Name	Description
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
out_prncp	Remaining outstanding principal for total amount funded. NOTE ONCE A LOAN IS REPAYED OR CHARGED OFF, THIS AMOUNT BECOMES 0.
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors. NOTE ONCE A LOAN IS REPAYED OR CHARGED OFF, THIS AMOUNT BECOMES 0.
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
total_rec_prncp	Principal received to date. NOTE THIS AMOUNT WILL SHOW WHETHER A BORROWER DID NOT REPAY IN FULL
total_rec_int	Interest received to date
total_rec_late_fee	Late fees received to date
recoveries	Post charge off gross recovery. NOTE IF A LOAN IS REPAYED, THIS AMOUNT IS 0.
collection_recovery_fee	Post charge off collection fee
last_pymnt_d	Last month payment was received
last_pymnt_amnt	Last total payment amount received
next_pymnt_d	Next scheduled payment date
last_credit_pull_d	The most recent month LC pulled credit for this loan
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
mths_since_last_major_derog	Months since most recent 90-day or worse rating
policy_code	Publicly available policy_code=1 / New products not publicly available policy_code=2
application_type	Indicates whether the loan is an individual application or a joint application with two coborrowers
annual_inc_joint	The combined self-reported annual income provided by the coborrowers during registration
dti_joint	A ratio calculated using the coborrowers total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the coborrowers combined self-reported monthly income
verification_status_joint	Indicates if income was verified by LC, not verified, or if the income source was verified
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
tot_coll_amt	Total collection amounts ever owed

Table 5.1: Description of the dataset variables as provided in the dataset downloaded from Kaggle (*continued*)

Variable Name	Description
tot_cur_bal	Total current balance of all accounts
open_acc_6m	Number of open trades in last 6 months
open_act_il	Number of currently active installment trades
open_il_12m	Number of installment accounts opened in past 12 months
open_il_24m	Number of installment accounts opened in past 24 months
mths_since_rcnt_il	Months since most recent instalment accounts opened
total_bal_il	Total current balance of all installment accounts
il_util	Ratio of total current balance to high credit/credit limit on all install acct
open(rv)_12m	Number of revolving trades opened in past 12 months
open(rv)_24m	Number of revolving trades opened in past 24 months
max_bal_bc	Maximum current balance owed on all revolving accounts
all_util	Balance to credit limit on all trades
total_rev_hi_lim	Total revolving high credit/credit limit
inq_fi	Number of personal finance inquiries
total_cu_tl	Number of finance trades
inq_last_12m	Number of credit inquiries in past 12 months
acc_open_past_24mths	Number of trades opened in past 24 months.
avg_cur_bal	Average current balance of all accounts
bc_open_to_buy	Total open to buy on revolving bankcards.
bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
chargeoff_within_12_mths	Number of charge-offs within 12 months
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
mo_sin_old_il_acct	Months since oldest bank instalment account opened
mo_sin_old_rev_tl_op	Months since oldest revolving account opened
mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened
mo_sin_rcnt_tl	Months since most recent account opened
mort_acc	Number of mortgage accounts.
mths_since_recent_bc	Months since most recent bankcard account opened.
mths_since_recent_bc_dlq	Months since most recent bankcard delinquency
mths_since_recent_inq	Months since most recent inquiry.
mths_since_recent_revol_delinq	Months since most recent revolving delinquency.
num_accts_ever_120_pd	Number of accounts ever 120 or more days past due
num_actv_bc_tl	Number of currently active bankcard accounts
num_actv_rev_tl	Number of currently active revolving trades

Table 5.1: Description of the dataset variables as provided in the dataset downloaded from Kaggle (*continued*)

Variable Name	Description
num_bc_sats	Number of satisfactory bankcard accounts
num_bc_tl	Number of bankcard accounts
num_il_tl	Number of installment accounts
num_op_rev_tl	Number of open revolving accounts
num_rev_accts	Number of revolving accounts
num_rev_tl_bal_gt_0	Number of revolving trades with balance >0
num_sats	Number of satisfactory accounts
num_tl_120dpd_2m	Number of accounts currently 120 days past due (updated in past 2 months)
num_tl_30dpd	Number of accounts currently 30 days past due (updated in past 2 months)
num_tl_90g_dpd_24m	Number of accounts 90 or more days past due in last 24 months
num_tl_op_past_12m	Number of accounts opened in past 12 months
pct_tl_nvr_dlq	Percent of trades never delinquent
percent_bc_gt_75	Percentage of all bankcard accounts > 75% of limit.
pub_rec_bankruptcies	Number of public record bankruptcies
tax_liens	Number of tax liens
tot_hi_cred_lim	Total high credit/credit limit
total_bal_ex_mort	Total credit balance excluding mortgage
total_bc_limit	Total bankcard high credit/credit limit
total_il_high_credit_limit	Total installment high credit/credit limit
revol_bal_joint	Total credit revolving balance
sec_app_earliest_cr_line	Earliest credit line at time of application for the secondary applicant. VARIABLE NOT USED. WE RELY ON THE MAIN BORROWER IN THE FIRST INSTANCE.
sec_app_inq_last_6mths	Credit inquiries in the last 6 months at time of application for the secondary applicant. VARIABLE NOT USED. WE RELY ON THE MAIN BORROWER IN THE FIRST INSTANCE.
sec_app_mort_acc	Number of mortgage accounts at time of application for the secondary applicant. VARIABLE NOT USED. WE RELY ON THE MAIN BORROWER IN THE FIRST INSTANCE.
sec_app_open_acc	Number of open trades at time of application for the secondary applicant. VARIABLE NOT USED. WE RELY ON THE MAIN BORROWER IN THE FIRST INSTANCE.
sec_app_revol_util	Ratio of total current balance to high credit/credit limit for all revolving accounts. VARIABLE NOT USED. WE RELY ON THE MAIN BORROWER IN THE FIRST INSTANCE.

Table 5.1: Description of the dataset variables as provided in the dataset downloaded from Kaggle (*continued*)

Variable Name	Description
sec_app_open_act_il	Number of currently active installment trades at time of application for the secondary applicant. VARIABLE NOT USED. WE RELY ON THE MAIN BORROWER IN THE FIRST INSTANCE.
sec_app_num_rev_accts	Number of revolving accounts at time of application for the secondary applicant. VARIABLE NOT USED. WE RELY ON THE MAIN BORROWER IN THE FIRST INSTANCE.
sec_app_chargeoff_within_12_mths	Number of charge-offs within last 12 months at time of application for the secondary applicant. VARIABLE NOT USED. WE RELY ON THE MAIN BORROWER IN THE FIRST INSTANCE.
sec_app_collections_12_mths_ex_med	Number of collections within last 12 months excluding medical collections at time of application for the secondary applicant. VARIABLE NOT USED. WE RELY ON THE MAIN BORROWER IN THE FIRST INSTANCE.
sec_app_mths_since_last_major_derog	Months since most recent 90-day or worse rating at time of application for the secondary applicant. VARIABLE NOT USED. WE RELY ON THE MAIN BORROWER IN THE FIRST INSTANCE.
hardship_flag	Flags whether or not the borrower is on a hardship plan
hardship_type	Describes the hardship plan offering
hardship_reason	Describes the reason the hardship plan was offered
hardship_status	Describes if the hardship plan is active, pending, cancelled, completed, or broken
deferral_term	Amount of months that the borrower is expected to pay less than the contractual monthly payment amount due to a hardship plan
hardship_amount	The interest payment that the borrower has committed to make each month while they are on a hardship plan
hardship_start_date	The start date of the hardship plan period
hardship_end_date	The end date of the hardship plan period
payment_plan_start_date	The day the first hardship plan payment is due. For example, if a borrower has a hardship plan period of 3 months, the start date is the start of the three-month period in which the borrower is allowed to make interest-only payments.
hardship_length	The number of months the borrower will make smaller payments than normally obligated due to a hardship plan

Table 5.1: Description of the dataset variables as provided in the dataset downloaded from Kaggle (*continued*)

Variable Name	Description
hardship_dpd	Account days past due as of the hardship plan start date
hardship_loan_status	Loan Status as of the hardship plan start date
orig_projected_additional_accrued_interest	The original projected additional interest amount that will accrue for the given hardship payment plan as of the Hardship Start Date. This field will be null if the borrower has broken their hardship payment plan.
hardship_payoff_balance_amount	The payoff balance amount as of the hardship plan start date
hardship_last_payment_amount	The last payment amount as of the hardship plan start date
disbursement_method	The method by which the borrower receives their loan. Possible values are: CASH, DIRECT_PAY
debt_settlement_flag	Flags whether or not the borrower, who has charged-off, is working with a debt-settlement company.
debt_settlement_flag_date	The most recent date that the Debt_Settlement_Flag has been set
settlement_status	The status of the borrower's settlement plan. Possible values are: COMPLETE, ACTIVE, BROKEN, CANCELLED, DENIED, DRAFT
settlement_date	The date that the borrower agrees to the settlement plan
settlement_amount	The loan amount that the borrower has agreed to settle for
settlement_percentage	The settlement amount as a percentage of the payoff balance amount on the loan
settlement_term	The number of months that the borrower will be on the settlement plan

5.4 Calculations of the internal rate of returns and month-to-default

The following shows two versions of the same code. The R version is provided because of the description of the assignment. However, the R version takes just under a full day to run. A Julia version, which is a direct translation of the R code, runs in about 150s (ca. 500x faster).

5.4.1 R code

```

1 ##########
2 #
3 # Given some numerical parameters describing a loan in the dataset, returns its Internal Rate
4 # of Return.
5 #
6 # In the first instance, the function creates a schedule of payments.
7 # In many cases, the schedule will be extremely simple: a series of 36 or 60 equal instalments.
8 #
9 # But in some cases, a loan repayment are accelerated. Therefore the total amount of interest will

```

```

10 # be lower than expected (but this is good for the investor because highe interest rate over
11 # shorter tenor.).
12 #
13 # In other cases, the borrower defaults. Overall payments are less than expected.
14 #
15 # Based on the limited information of the dataset, the function makes educated guesses on the exact ..
16 #
17 # WARNING: THIS IS NOT OPTIMISED. RUNNING THIS FOR ALL LOANS (1.3 MLN OF THEM) TAKES CA.20 HOURS !!!
18 #
19 calculateIRR <- function(loanNumber = 1,
20                         loan,
21                         intRate,
22                         term = 36,
23                         totalPaid,
24                         totalPrincipalPaid,
25                         totalInterestPaid,
26                         recoveries = 0,
27                         lateFees = 0,
28                         showSchedule = FALSE) {
29   require(tidyverse)
30
31   # number of monthly payments.
32   # It exceeds 60 months in case recoveries on a 60-month loan takes the schedule after 60 months.
33   nMonths <- 90
34
35   # Months after which a loan defaults (normal tenor if no default or early prepayment)
36   monthDefault = term
37
38   # Note: *100 /100 to calculate in cent because ceiling cannot specify significant digits.
39   installment <-
40     ceiling(100 * loan * intRate / 12 / (1 - 1 / (1 + intRate / 12) ^ term)) / 100
41
42   # We create a schedule
43   schedule <- tibble(
44     month = 0:nMonths,
45     monthlyPayment = 0.0,
46     totalPandI = 0.0,
47     totalI = 0.0,
48     totalP = 0.0
49   )
50
51   for (i in 2:(nMonths + 2)) {
52     # Get situation at the end of previous month
53     previousTotalPandI <- as.numeric(schedule[i - 1, "totalPandI"])
54     previousTotalP    <- as.numeric(schedule[i - 1, "totalP"])
55     previousTotalI    <- as.numeric(schedule[i - 1, "totalI"])
56
57     # This is the beginning of a new month. First and foremost, the borrower is expected to pay the
58     # accrued interest on amount of principal outstanding.
59     # ceiling doesn't seem accept to accept significative digits.
60     accruedInterest <-
61       ceiling(100 * (loan - previousTotalP) * intRate / 12) / 100
62
63     # If that amount takes the schedule above the total amount of interest shown in the data set,
64     # we should stop the schedule at this point
65     if (previousTotalI + accruedInterest > totalInterestPaid) {
66       # We stop the normal schedule at this date.
67       # Interest is paid (although less than scheduled)

```

```

68     schedule[i, "monthlyPayment"] <-
69         totalInterestPaid - previousTotalI
70
71     # As well as whatever principal is left as per the dataset
72     schedule[i, "monthlyPayment"] <-
73         schedule[i, "monthlyPayment"] + totalPrincipalPaid - previousTotalP
74
75     # Then 3-month after the last payment date, recoveries and and later fees are paid
76     schedule[i + 3, "monthlyPayment"] <-
77         schedule[i + 3, "monthlyPayment"] + recoveries + lateFees
78
79     # Not really useful, but for completeness
80     schedule[i, "totalPandI"] <- totalPaid
81     schedule[i, "totalI"]      <- totalInterestPaid
82     schedule[i, "totalP"]      <- totalPrincipalPaid
83
84     # If total principal paid is less than borrower, then it is a default, and the monthDefault
85     # is adjusted.
86     if (totalPrincipalPaid < loan) {
87         monthDefault = i
88     }
89
90     # No more payments to add to the schedule
91     break()
92
93 } else {
94     # Deal with normal schedule
95     schedule[i, "monthlyPayment"] <- installment
96     schedule[i, "totalPandI"] <-
97         schedule[i - 1, "totalPandI"] + installment
98     schedule[i, "totalI"]      <-
99         schedule[i - 1, "totalI"]   + accruedInterest
100    schedule[i, "totalP"]      <-
101        schedule[i - 1, "totalP"] + installment - accruedInterest
102    }
103 }
104
105 # At this point schedule[, "monthlyPayment"] contains the schedule of all payments, but needs to
106 # include the initial loan.
107 schedule[1, "monthlyPayment"] <- -loan
108
109 if (showSchedule) {
110     schedule %>% view()
111 }
112
113 NPV <- function(interest, cashFlow) {
114     t = 0:(length(cashFlow) - 1)
115     sum(cashFlow / (1 + interest) ^ t)
116 }
117
118 IRR <- function(CF) {
119     res <- NA
120     try({
121         res <- uniroot(NPV, c(-0.9, 1), cashFlow = CF)$root
122     },
123         silent = TRUE)
124     return(res)
125 }
```

```

126
127     return(tibble(
128         loanID = loanNumber,
129         IRR = round(as.numeric(IRR(
130             schedule$monthlyPayment
131         ) * 12), digits = 4),
132         monthDefault = monthDefault
133     ))
134 }
135
136 loanNumberIRR <- function(loanNumber) {
137     require(tidyverse)
138
139     l <- loans %>% filter(loanID == loanNumber)
140     calculateIRR(
141         loanNumber = l$loanID,
142         loan = l$funded_amnt, intRate = l$int_rate, term = l$term,
143         totalPaid = l$total_pymnt, totalPrincipalPaid = l$total_rec_prncp, totalInterestPaid = l$total_r
144         recoveries = l$recoveries, lateFees = l$total_rec_late_fee,
145         showSchedule = TRUE
146     )
147 }
1
2 # Calculate all the IRRs and month of default for all the loans.
3 # WARNING: This takes around a full day to run!!!!
4 #
5 # The actual data was generated by the Julia version, with cross-checks.
6 # Julia version takes about 150 sec on the same unoptimised code.
7 #
8 local({
9     loansIRR <-
10     loans %>%
11     rowwise() %>%
12     do(calculateIRR(loanNumber = .$loanID,
13                     loan = .$funded_amnt, intRate = .$int_rate, term = .$term,
14                     totalPaid = .$total_pymnt, totalPrincipalPaid = .$total_rec_prncp,
15                     totalInterestPaid = .$total_rec_int,
16                     recoveries = .$recoveries, lateFees = .$total_rec_late_fee))
17
18     saveRDS(loansIRR, "datasets/lending_club_IRRs.rds")
19
20 })

```

5.4.2 Julia code

5.4.2.1 Internal Rate of Return

```

1 #####
2 ##
3 ## Prepare datasets
4 ##
5
6 ## Previously saved from R with:
7 ##     lending_club <- readRDS("lending_club.rds"); write.csv(lending_club, "lending_club.rds")
8 ##
9 ## WARNING: 1.7GB on disk
10 ##

```

```

11  using CSV
12  lendingClub = CSV.read("datasets/lending_club.csv"; delim = ",")
13
14
15
16 #####
17 ##
18 ## IRR calculations
19 ##
20 ## Given some numerical parameters describing a loan in the dataset, returns its Internal Rate
21 ## of Return.
22 ##
23 ## In the first instance, the function creates a schedule of payments.
24 ## In many cases, the schedule will be extremely simple: a series of 36 or 60 equal instalments.
25 ##
26 ## But in some cases, a loan repayment are accelerated. Therefore the total amount of interest will
27 ## be lower than expected (but this is good for the investor because highe interest rate over
28 ## shorter tenor.).
29 ##
30 ## In other cases, the borrower defaults. Overall payments are less than expected.
31 ##
32 ## Based on the limited information of the dataset, the function makes educated guesses on the exact
33 ## schedule.
34 ##
35 using DataFrames, Roots
36
37 function calculateIRR(; loanNumber = 1, loan = 0.0, intRate = 0.0, term = 36,
38   totalPaid = 0.0, totalPrincipalPaid = 0.0, totalInterestPaid = 0.0,
39   recoveries = 0.0, lateFees = 0.0,
40   showSchedule = false)
41
42   # number of monthly payments.
43   # It exceeds 60 months in case recoveries on a 60-month loan takes the schedule after 60 months.
44   nMonths = 90
45
46   # Months after which a loan defaults (normal tenor if no default or early prepayment)
47   monthDefault = term
48
49   # Note: *100 /100 to calculate in cent because ceiling cannot specify significant digits.
50   installment = ceil(loan * intRate / 12 / (1 - 1 / (1 + intRate / 12) ^ term), digits = 2)
51
52   # We create a schedule
53   schedule = DataFrame(month = 0:nMonths, monthlyPayment = 0.0,
54                         totalPandI = 0.0, totalI = 0.0, totalP = 0.0)
55
56   for i in 2:(nMonths + 1)
57     # Get situation at the end of previous month
58     previousTotalPandI = schedule[i - 1, :totalPandI]
59     previousTotalP      = schedule[i - 1, :totalP]
60     previousTotalI      = schedule[i - 1, :totalI]
61
62     # This is the beginning of a new month. First and foremost, the borrower is expected to pay the
63     # accrued interest on amount of principal outstanding.
64     # The installment is expected to cover that amount of interest and the rest goes to
65     # reducing the principal due outstanding.
66     accruedInterest = ceil((loan - previousTotalP) * intRate / 12; digits = 2)
67     decreasePrincipal = installment - accruedInterest
68

```

```

69  # If that amount takes the schedule above the total amount of interest shown in the data set,
70  # we should stop the schedule at this point
71  # This is a shortcut since we could have a payment higher than the interest due, but not enough
72  # to cover the expected principal repayment. However, it works well in practice.
73  if previousTotalI + accruedInterest > totalInterestPaid
74
75  # We stop the normal schedule at this date.
76  # Interest is paid (although less than scheduled)
77  schedule[i, :monthlyPayment] = totalInterestPaid - previousTotalI
78
79  # As well as whatever principal is left as per the dataset
80  schedule[i, :monthlyPayment] = schedule[i, :monthlyPayment] + totalPrincipalPaid - previousTotalI
81
82  # Then 3-month after the last payment date, recoveries and late fees are paid
83  schedule[i + 3, :monthlyPayment] = schedule[i + 3, :monthlyPayment] + recoveries + lateFees
84
85  # Not really useful, but for completeness
86  schedule[i, :totalPandI] = totalPaid
87  schedule[i, :totalI]     = totalInterestPaid
88  schedule[i, :totalP]    = totalPrincipalPaid
89
90  # If total principal paid is less than borrower, then it is a default, and the monthDefault
91  # is adjusted.
92  if (totalPrincipalPaid < loan)
93      monthDefault = i
94  end
95
96  # No more payments to add to the schedule
97  break
98
99 else
100  # Deal with normal schedule
101  schedule[i, :monthlyPayment] = installment
102  schedule[i, :totalPandI]     = schedule[i - 1, :totalPandI] + installment
103  schedule[i, :totalI]        = schedule[i - 1, :totalI]     + accruedInterest
104  schedule[i, :totalP]        = schedule[i - 1, :totalP]     + installment - accruedInterest
105 end
106 end
107
108 # At this point schedule[, :monthlyPayment] contains the schedule of all payments, but needs to
109 # include the initial loan.
110 schedule[1, :monthlyPayment] = -loan
111
112 if (showSchedule)
113     print(schedule)
114 end
115
116 cashFlow = schedule[:, :monthlyPayment]
117
118 ##
119 ## Finding the IRR is equivalent to finding the root such that the NPV of the cash flow is zero.
120 ## Julia has a function (see below) called `find_zero` to do that which requires a function to
121 ## be zeroed. This helper function is defined as NPV.
122 ##
123 function NPV(interest)
124     t = 0:(length(cashFlow) - 1)
125
126     ## If you are new to Julia, note the dot before the operation. This indicates that the

```

```

127     ## operation has to be done element-wise (called `broadcasting` in Julia-ese).
128     ## Otherwise, Julia would try to divide one vector by another vector, which makes no sense.
129     ## This is also exactly the approach taken in Matlab/Octave.
130     return sum(cashFlow ./ (1 + interest) .^ t)
131 end
132
133     ## Finds the root, catching any problems which would instead return the R equivalent of NA
134 rootInterest = try
135         round(12 * find_zero(NPV, (-0.9, 1.0), Bisection(); xatol = 0.000001); digits = 4)
136     catch e
137         NaN
138     end
139
140     return(
141         loanID = loanNumber,
142         IRR = rootInterest,
143         monthDefault = monthDefault
144     ))
145 end
146
147
148     ##
149     ## Calculate the IRR and repayment schedule of a particular loan identified by its loanID
150 function loanNumberIRR(loanNumber)
151     l = lc[ lc[:, :Column1] .== loanNumber, :]
152     global lc
153     calculateIRR(loanNumber = l[1, :Column1],
154                 loan = l[1, :funded_amnt], intRate = l[1, :int_rate], term = l[1, :tenor],
155                 totalPaid = l[1, :total_pymnt], totalPrincipalPaid = l[1, :total_rec_prncp],
156                 totalInterestPaid = l[1, :total_rec_int],
157                 recoveries = l[1, :recoveries], lateFees = l[1, :total_rec_late_fee],
158                 showSchedule = true)
159 end
160
161
162 #####
163 ##
164 ## Quick check
165 ##
166 calculateIRR(loanNumber = 1, loan = 5600, intRate = 0.1299, term = 36,
167                 totalPaid = 6791.72, totalPrincipalPaid = 5600, totalInterestPaid = 1191.72,
168                 recoveries = 0, lateFees = 0,
169                 showSchedule = true)
170
171 calculateIRR(loanNumber = 1, loan = 35000, intRate = 0.1820, term = 60,
172                 totalPaid = 26600.1, totalPrincipalPaid = 3874.72, totalInterestPaid = 5225.38,
173                 recoveries = 17500, lateFees = 0.0,
174                 showSchedule = false)
175
176 calculateIRR(loanNumber = 1734666, loan = 35000, intRate = 0.0797, term = 36,
177                 totalPaid = 1057.04, totalPrincipalPaid = 863.83, totalInterestPaid = 193.72,
178                 recoveries = 0, lateFees = 0,
179                 showSchedule = false)
180
181
182
183
184     ##

```

```

185 ## Look for the loans which have gone to their end
186 ##
187 indextmp = (lendingClub.loan_status == "Fully Paid") .|
188     (lendingClub.loan_status == "Charged Off") .|
189     (lendingClub.loan_status == "Does not meet the credit policy. Status:Charged Off") .|
190     (lendingClub.loan_status == "Does not meet the credit policy. Status:Fully Paid")
191
192 ## Create the dataset we will use - Should be the same as lending_club_reformatted_paid.rds
193 lc = lendingClub[indextmp, :]
194
195 ## Select relevant variables to calculate profitability
196 ## Column1 contains the loanID's
197 cols = [:Column1, :funded_amnt, :int_rate, :term,
198     :total_pymnt, :total_rec_prncp, :total_rec_int,
199     :recoveries, :total_rec_late_fee]
200
201 lc = select(lc, cols)
202
203 ## Interest rates as percentage
204 lc[:, :int_rate] = lc[:, :int_rate] ./ 100
205
206 ## Create a new column
207 lc[:tenor] = 0
208
209 ## that will record the official loan tenor as a number (instead of string)
210 lc[startswith.( lc[:, :term], " 36"), :tenor] .= 36
211 lc[startswith.( lc[:, :term], " 60"), :tenor] .= 60
212
213 ## New data frame to store the results
214 IRR_Result = DataFrame(loanID = zeros(Int64, nrow(lc)),
215                         IRR = zeros(Float64, nrow(lc)),
216                         monthDefault = zeros(Int64, nrow(lc)))
217
218
219 # ~150 sec. to do the whole dataset
220 @time for i in 1:nrow(lc)
221     global IRR_Result
222
223     # Use multiple-return-value
224     (IRR_Result[i, :loanID], IRR_Result[i, :IRR], IRR_Result[i, :monthDefault]) =
225         calculateIRR(
226             loanNumber = lc[i, :Column1],
227             loan = lc[i, :funded_amnt], intRate = lc[i, :int_rate], term = lc[i, :tenor],
228             totalPaid = lc[i, :total_pymnt], totalPrincipalPaid = lc[i, :total_rec_prncp],
229             totalInterestPaid = lc[i, :total_rec_int],
230             recoveries = lc[i, :recoveries], lateFees = lc[i, :total_rec_late_fee],
231             showSchedule = false)
232 end
233
234
235 IRR_Result[1:10,:]
236 # Check
237 loanNumberIRR(171)
238
239 CSV.write("datasets/loanIRR.csv", IRR_Result)

```

5.4.2.2 Credit margins

```
1 #####  
2 ##  
3 ## Prepare datasets  
4 ##  
5 ##  
6 ## Previously saved from R with:  
7 ##     lending_club <- readRDS("lending_club.rds"); write.csv(lending_club, "lending_club.rds")  
8 ##  
9 ## WARNING: 1.7GB on disk  
10##  
11using CSV  
12lendingClub = CSV.read("datasets/lending_club.csv"; delim = ",")  
13  
14  
15#####  
16##  
17##  
18## CREDIT MARGIN  
19##  
20## This method of approximating the credit margin is far less sophisticated than what FI's do.  
21##  
22## We need to calculate what the credit margin should be on a defaulted loan to get a nil NPV.  
23##  
24## On a risk free loan the CF will be P+I at risk-free on _both_ borrowing and lending sides.  
25##  
26## On a defaulted loan, the CF will be:  
27##     borrowing unchanged = P&I at risk-free  
28##     and  
29##     lending P&I at (risk-free + credit margin) until before default, then recoveries+fees.  
30##  
31## The principal amortisation profile depends on the credit margin used. We will arbitrarily use  
32## 20% which is a conservative assumption.  
33##  
34## credit risk on that CF should be nil with the right margin when discounted at risk-free.  
35##  
36## The function is very similar to the IRR calculation.  
37##  
38  
39using DataFrames, Roots  
40  
41# number of monthly payments to model  
42# It exceeds 60 months in case recoveries on a 60-month loan takes the schedule after 60 months.  
43const nMonths = 90  
44  
45  
46  
47# Sculpt credit foncier profiles over 36 and 60 months at 20% per annum.  
48# The profile is expressed as percentage of loan amount  
49function CreateCreditFoncier(;n = 36, riskFree = 0.0)  
50  
51    instalment = 1 * riskFree/12 * 1 / (1 - 1 / (1 + riskFree/12) ^ n)  
52  
53    # We create a schedule  
54    schedule = DataFrame(month = 0:nMonths, payment = 0.0)  
55  
56    # Add the day 1 principal outlay
```

```

57     schedule[1, :payment] = -1
58     schedule[2:(n+1), :payment] = instalment
59
60     return(schedule)
61 end
62
63
64 # Solve for the credit margin
65 function CreditMargin(; loanNumber = 1, loan = 1000.0, intRate = 0.05, term = 36,
66     totalPaid = 1000.0, totalPrincipalPaid = 700.0, totalInterestPaid = 50.0,
67     recoveries = 0.0, lateFees = 0.0,
68     riskFree = 0.01,
69     showSchedule = false)
70
71 # Months after which a loan defaults (normal tenor if no default or early prepayment)
72 monthDefault = term
73
74 # Monthly instalment
75 instalment = ceil(loan * intRate/12 / (1 - 1 / (1 + intRate/12) ^ term), digits = 2)
76
77 # Create a blank schedule
78 schedule = DataFrame(month = 0:nMonths, monthlyPayment = 0.0,
79                     principalPayment = 0.0,
80                     totalPandI = 0.0, totalI = 0.0, totalP = 0.0)
81
82 for i in 2:(nMonths + 1)
83     # Get situation at the end of previous month
84     previousTotalPandI = schedule[i - 1, :totalPandI]
85     previousTotalP    = schedule[i - 1, :totalP]
86     previousTotalI   = schedule[i - 1, :totalI]
87
88     # This is the beginning of a new month. First and foremost, the borrower is expected to pay the
89     # accrued interest on amount of principal outstanding.
90     # The instalment is expected to cover that amount of interest and the rest goes to
91     # reducing the principal due outstanding.
92     accruedInterest = ceil((loan - previousTotalP) * intRate/12; digits = 2)
93     decreasePrincipal = instalment - accruedInterest
94
95     # If that amount takes the schedule above the total amount of interest shown in the data set,
96     # we should stop the schedule at this point
97     # This is a shortcut since we could have a payment higher than the interest due, but not enough
98     # to cover the expected principal repayment. However, it works well in practice.
99     if previousTotalI + accruedInterest > totalInterestPaid
100
101     # We stop the normal schedule at this date.
102     # Interest is paid (although less than scheduled)
103     schedule[i, :monthlyPayment] = totalInterestPaid - previousTotalI
104
105     # Whatever principal is left as per the dataset
106     schedule[i, :monthlyPayment] += totalPrincipalPaid - previousTotalP
107
108     # Then 3-month after the last payment date, recoveries and late fees are paid
109     schedule[i + 3, :principalPayment] += recoveries + lateFees
110
111     # Not really useful, but for completeness
112     schedule[i, :totalPandI] = totalPaid
113     schedule[i, :totalI]     = totalInterestPaid
114     schedule[i, :totalP]     = totalPrincipalPaid

```

```

115
116      # If total principal paid is less than borrower, then it is a default, and the monthDefault
117      # is adjusted.
118      if (totalPrincipalPaid < loan)
119          monthDefault = i
120      end
121
122      # No more payments to add to the schedule
123      break
124
125  else
126      # Deal with normal schedule
127      schedule[i, :monthlyPayment] = instalment
128      schedule[i, :principalPayment] = decreasePrincipal
129      schedule[i, :totalPandI] = schedule[i-1, :totalPandI] + instalment
130      schedule[i, :totalI] = schedule[i-1, :totalI] + accruedInterest
131      schedule[i, :totalP] = schedule[i-1, :totalP] + decreasePrincipal
132  end
133 end
134
135 # At this point schedule[, :monthlyPayment] contains the schedule of all payments, but needs to
136 # include the initial loan.
137 schedule[1, :principalPayment] = -loan
138
139 if (showSchedule)
140     println("Payments")
141     println(schedule)
142 end
143
144 # Principal profile on the borrowing side
145 creditFoncier = round.(CreateCreditFoncier(n = term, riskFree = riskFree)[:, :payment] .* loan;
146 digits = 2)
147
148 # For a given margin, calculate the _net_ NPV between the what is borrowed at the risk-free rate
149 # and what is earned on the loan principal profile (possibly shortned because of default)
150 # carrying an interest of risk-free + credit margin
151 function NetNPV(margin)
152     # We need to store the calculated interest
153     interestSchedule = DataFrame(month = 0:nMonths, interestPayment = 0.0)
154
155     # For each month, calculate the amount of interest with the credit margin
156     for i in 2:(monthDefault + 1)
157         outstandingPrincipal = sum(schedule[1:(i - 1), :principalPayment])
158         interestSchedule[i, :interestPayment] =
159             -round((riskFree + margin)/12 * outstandingPrincipal; digits = 2)
160     end
161
162     # Net final cashflow is:
163     #   total principal and interest cashflow on the lending side
164     # less
165     #   borrowing profile
166     cashFlow = schedule[:, :principalPayment] .+ interestSchedule[:, :interestPayment]
167     cashFlow = cashFlow .- creditFoncier
168
169     return sum(cashFlow ./ (1 + riskFree/12) .^ (0:nMonths))
170 end
171
172 # rootInterest = round(find_zero(NetNPV, (-0.5, 10), Bisection()); digits = 6)

```

```

173     rootInterest = try
174         rootInterest = round(find_zero(NetNPV, (-0.5, 10), Bisection()); digits = 6)
175     catch e
176         NaN
177     end
178
179
180     return(
181         loanID = loanNumber,
182         creditMargin = rootInterest,
183         monthDefault = monthDefault
184     )))
185 end
186
187
188 #####
189 ##
190 ## Quick check
191 ##
192 CreditMargin(loanNumber = 1, loan = 5600, intRate = 0.1299, term = 36,
193                 totalPaid = 6791.72, totalPrincipalPaid = 5600, totalInterestPaid = 1191.72,
194                 recoveries = 0, lateFees = 0,
195                 riskFree = 0.02, showSchedule = false)
196
197 CreditMargin(loanNumber = 1, loan = 35000, intRate = 0.1820, term = 60,
198                 totalPaid = 26600.1, totalPrincipalPaid = 3874.72, totalInterestPaid = 5225.38,
199                 recoveries = 17500, lateFees = 0.0,
200                 riskFree = 0.02, showSchedule = true)
201
202 CreditMargin(loanNumber = 1734666, loan = 35000, intRate = 0.0797, term = 36,
203                 totalPaid = 1057.04, totalPrincipalPaid = 863.83, totalInterestPaid = 193.72,
204                 recoveries = 0, lateFees = 0,
205                 riskFree = 0.02, showSchedule = true)
206
207
208 ##
209 ## Look for the loans which have gone to their end
210 ##
211 indextmp = (lendingClub.loan_status == "Fully Paid") .|
212     (lendingClub.loan_status == "Charged Off") .|
213     (lendingClub.loan_status == "Does not meet the credit policy. Status:Charged Off") .|
214     (lendingClub.loan_status == "Does not meet the credit policy. Status:Fully Paid")
215
216 ## Create the dataset we will use - Should be the same as lending_club_reformatted_paid.rds
217 lc = lendingClub[indextmp, :]
218
219 ## Select relevant variables to calculate profitability
220 ## Column1 contains the loanID's
221 cols = [:Column1, :funded_amnt, :int_rate, :term,
222           :total_pymnt, :total_rec_prncp, :total_rec_int,
223           :recoveries, :total_rec_late_fee]
224
225 lc = select(lc, cols)
226
227 ## Interest rates as percentage
228 lc[!, :int_rate] = lc[!, :int_rate] ./ 100
229
230 ## Create a new column

```

```

231 lc[:tenor] = 0
232
233 ## that will record the official loan tenor as a number (instead of string)
234 lc[startswith.( lc[!, :term], " 36"), :tenor] .= 36
235 lc[startswith.( lc[!, :term], " 60"), :tenor] .= 60
236
237 ## New data frame to store the results
238 creditMargin_Result = DataFrame(loanID = zeros(Int64, nrow(lc)),
239                                 creditMargin = zeros(Float64, nrow(lc)),
240                                 monthDefault = zeros(Int64, nrow(lc)))
241
242
243 # ~4,700 sec. to do the whole dataset
244 @time for i in 1:nrow(lc)
245     global creditMargin_Result
246
247     # Use multiple-return-value
248     # 1h17m runtime
249     (creditMargin_Result[i, :loanID],
250      creditMargin_Result[i, :creditMargin],
251      creditMargin_Result[i, :monthDefault]) =
252         CreditMargin(
253             loanNumber = lc[i, :Column1],
254             loan = lc[i, :funded_amnt], intRate = lc[i, :int_rate], term =lc[i, :tenor],
255             totalPaid = lc[i, :total_pymnt], totalPrincipalPaid = lc[i, :total_rec_prncp],
256             totalInterestPaid = lc[i, :total_rec_int],
257             recoveries = lc[i, :recoveries], lateFees = lc[i, :total_rec_late_fee],
258             riskFree = 0.02,
259             showSchedule = false)
260 end
261
262
263 creditMargin_Result[1:10,:]
264
265 CSV.write("datasets/CreditMargins.csv", creditMargin_Result)

```

5.4.3 Maxima derivation of the cost function

```

PDF1(x, Q) := alpha1( Q) * sqrt( 1 / ( 2 * pi)) *
              exp( - 1 / 2*(( log( -( x - m1( Q)) / m1( Q)) + sigma1( Q) ^ 2) / sigma1( Q))
                  ( -( x - m1(Q)) * sigma1( Q)) ;

PDF2(x, Q) := alpha2( Q) * sqrt( 1 / ( 2 * pi)) *
              exp( - 1 / 2*(( log( -( x - m2( Q)) / m2( Q)) + sigma2( Q) ^ 2) / sigma2( Q))
                  ( -( x - m2( Q)) * sigma2( Q)) ;

PDF3(x, Q) := alpha3( Q) * sqrt( 1 / ( 2 (* pi)) *
              exp( - 1 / 2*(( log( -( x - m3( Q)) / m3( Q)) + sigma3( Q) ^ 2) / sigma3( Q))
                  ( -( x - m3( Q)) * sigma3( Q)) ;

PDF4(x, Q) := alpha4( Q) * sqrt( 1 / ( 2 * pi)) *
              exp( - 1 / 2*(( log( ( x - m4( Q)) / m4( Q)) + sigma4( Q) ^ 2) / sigma4( Q))
                  ( x - m4( Q)) * sigma4( Q)) ;

```

```

alpha1(Q) := am1* Q + an1 ;
alpha2(Q) := am2* Q + an2 ;
alpha3(Q) := am3* Q + an3 ;
alpha4(Q) := am4* Q + an4 ;

m1(Q) := mm1* Q + mn1 ;
m2(Q) := mm2* Q + mn2 ;
m3(Q) := mm3* Q + mn3 ;
m4(Q) := mm4* Q + mn4 ;

sigma1(Q) := sm1* Q + sn1 ;
sigma2(Q) := sm2* Q + sn2 ;
sigma3(Q) := sm3* Q + sn3 ;
sigma4(Q) := sm4* Q + sn4 ;

J(x, Q): = -( x - ( PDF1( x, Q) + PDF2( x, Q) + PDF3( x, Q) + PDF4( x, Q))) ^ 2 ;

diff( PDF1(x, Q), Q) ;

```

5.5 System version

1 ##	sysname	release
2 ##	"Linux"	"5.3.0-21-generic"
3 ##	version	nodename
4 ## "#22-Ubuntu SMP Tue Oct 29 22:55:51 UTC 2019"		"x260"
5 ##	machine	login
6 ##	"x86_64"	"unknown"
7 ##	user	effective_user
8 ##	"emmanuel"	"emmanuel"

Bibliography

- California, L. C. S. F. (2019). Prospectus Regulatory Filing S3-ASR for Member Payment Dependent Notes. <https://www.sec.gov/Archives/edgar/data/1409970/000140997019000988/0001409970-19-000988-index.htm>. [Note: Accessed 31 October 2019].
- Kan, L. C. P. W. (2019). Kaggle - LendingClub dataset. <https://www.kaggle.com/wendykan/lending-club-loan-data>. [Note: Accessed 31 August 2019].
- Kim, A. and Cho, S.-B. (2019). An ensemble semi-supervised learning method for predicting defaults in social lending. *Engineering Applications of Artificial Intelligence*, 81:193–199.
- Peng, R. (2012). *Exploratory data analysis with R*. Lulu. com.
- Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.