

LendingClub Loans Pricing

HarvardX - PH125.9x Data Science Capstone

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

December 07, 2019

Contents

TODO	5
0.1 Dataset	5
0.2 Model preparation	5
Introduction	6
1 Internal Rate of Return, Credit Margins and Net Present Values	8
1.1 Important warning	8
1.2 Background	8
1.3 Internal Rate of Return	9
1.4 Dataset calculation	10
1.4.1 IRR	10
1.4.2 Credit Margins	11
1.4.3 NPV	11
2 Dataset	12
2.1 Preamble	12
2.2 General presentation	12
2.2.1 Business volume	12
2.2.2 Loan lifecycle and status	14
2.2.3 Loan application	15
2.3 Rates	15
2.3.1 IRR and required credit margins	15
2.3.2 Dataset	16
2.3.3 Interest rates	16
2.3.4 Purpose	19
2.3.5 Payments	21
2.4 Net present value	21
2.4.1 Average NPV and credit margin by subgrade	22
2.4.2 Principal losses	22
2.4.3 Distribution of principal losses by rating	22
2.4.4 NPV distribution by rating	22
2.5 Variables	24
2.5.1 Identification	26
2.6 Loan decision	26
2.7 Payment-related information	27
3 Logistic Regression Model and Credit Scorecard	28
3.1 Introduction	28
3.1.1 Split	28
3.1.2 The modeling task	28

3.2	Logistic Regression	29
3.2.1	Data preparation	29
3.3	Binning and Weight of Evidence	30
3.3.1	Loop through all variables	31
3.3.2	Select relevant variables	33
3.3.3	Create data table with only binary variables (transform every bin to a 0/1 value)	35
3.3.4	Comparison of individual characteristics	36
3.4	Logistic Regression	40
3.4.1	Logistic regression using the SpeedGLM package	40
3.4.2	Remove identical bins	41
3.4.3	First training on variables previously selected on their Information Value	42
3.4.4	Second training	44
3.4.5	Third training	45
3.5	Model result	46
3.5.1	Final list of variables	46
3.5.2	Scoring	46
3.6	Training set	47
3.6.1	Histogram of the scores	48
3.7	Test set	49
3.8	Correlation matrix	51
3.9	ROC Curve	51
3.10	Bin test set by prediction	51
4	Conclusion	52
5	Stochastic Gradient Descent	53
5.1	Description of the model	53
5.2	Gradient descent	53
5.2.1	Generalities	53
5.3	Stochastic Gradient Descent	54
5.3.1	Stochastic Gradient Descent (SGD)	56
5.4	Various errands	60
5.4.1	Optimising the scorecard (or rating) depending on the NPV curve	60
Appendix		62
5.5	List of assumptions / limitations regarding the dataset	62
5.6	Data preparation and formatting	62
5.6.1	LendinClub dataset	63
5.6.2	Zip codes and FIPS codes	68
5.6.3	Market interest rates	68
5.6.4	Macro-economical data	70
5.7	List of variables	72
5.8	Calculations of the internal rate of returns and month-to-default	78
5.8.1	R code	78
5.8.2	Julia code	81
5.8.3	Maxima derivation of the cost function	81
5.9	System version	82

List of Tables

2.1	Number of loans per status	13
2.2	Number of loans per purpose	20
2.3	Matured loans per status	26
5.1	Description of the dataset variables as provided in the dataset downloaded from Kaggle	72

List of Figures

2.1	Business volume written per month	13
2.2	Credit margins per grade over time	16
2.3	Credit margins per grade over time	17
2.4	Credit margins per grade over time	17
2.5	Interest rates given rating	18
2.6	Interest rate per grade over time	18
2.7	Historical Swap Rates	19
2.8	Credit margins per grade over time	19
2.9	Histograms of credit margins per purpose	20
2.10	Boxplots of credit margins per purpose	21
2.11	Average NPV et credit margin (%) depending on sub-rating	22
2.12	Distribution of NPV (%) depending on rating (y-axis square-root scaling)	23
2.13	Distribution of NPV (%) depending on rating (y-axis square-root scaling)	23
2.14	NPV % over 120% (no y-axis scaling)	24
2.15	NPV % around 41% (no y-axis scaling)	24
2.16	NPV % around -1% (no y-axis scaling)	25
2.17	NPV % for close to total loss % (no y-axis scaling)	25
2.18	Funding and Write-offs by Sub-grades	26
5.1	Scikit Learn algorithm cheat-sheet	53

TODO

0.1 Dataset

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

0.2 Model preparation

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

Correlation matrix?

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

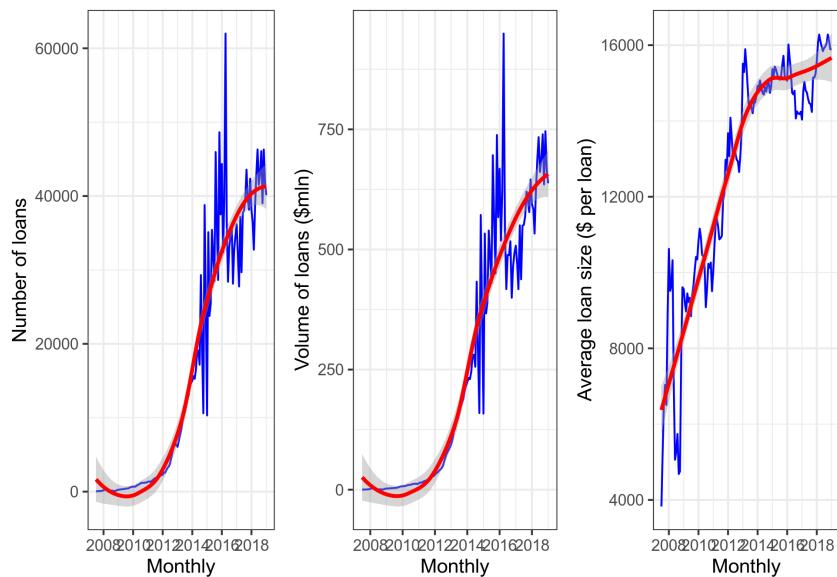


Figure 2.1: Business volume written per month

" 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>

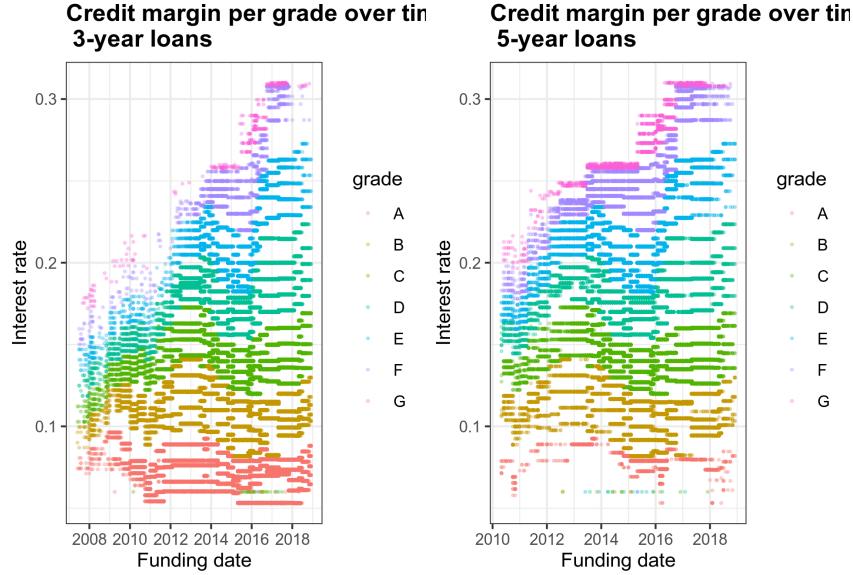


Figure 2.2: Credit margins per grade over time

- 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.

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.

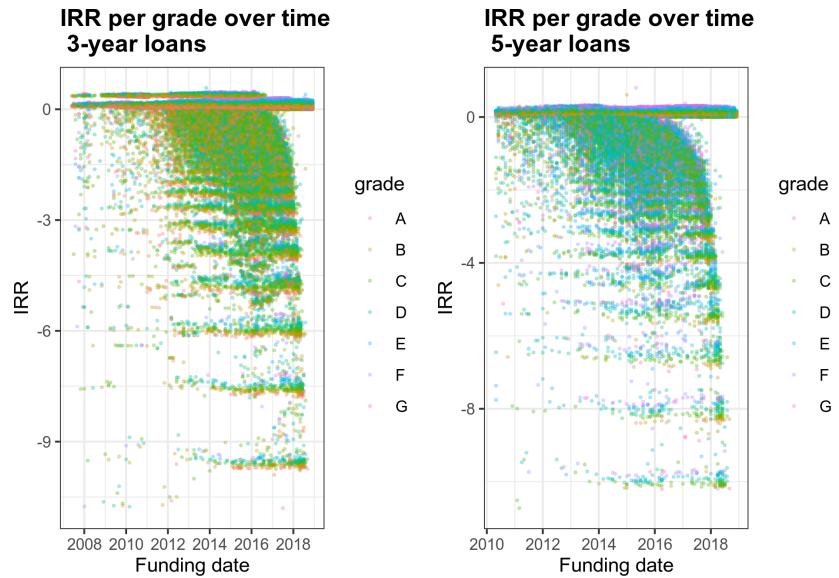


Figure 2.3: Credit margins per grade over time

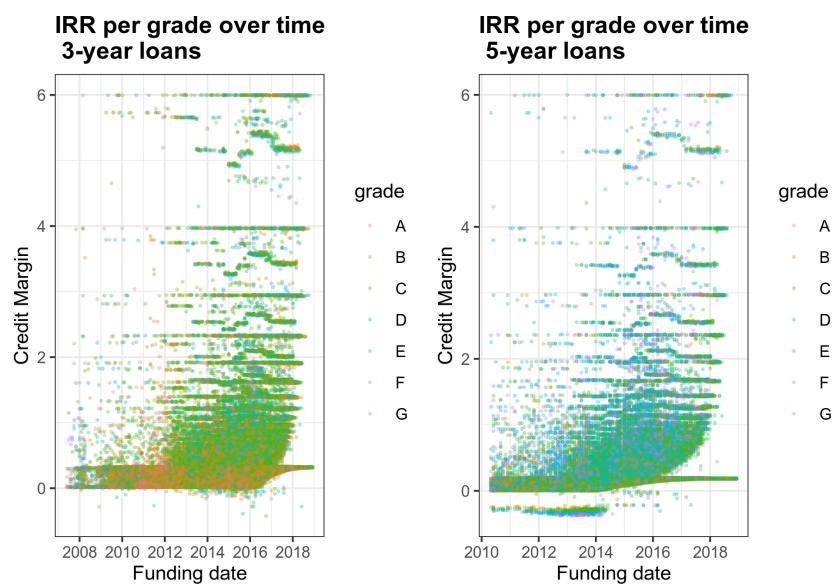


Figure 2.4: Credit margins per grade over time

Figure 2.5⁸ shows the predetermined interest rate depending on the initial rating as of July 2019.

A1	6.46%	B1	10.33%	C1	14.30%	D1	18.62%
A2	7.02%	B2	11.02%	C2	15.24%	D2	20.55%
A3	7.56%	B3	11.71%	C3	16.12%	D3	23.05%
A4	8.19%	B4	12.40%	C4	16.95%	D4	25.65%
A5	8.81%	B5	13.08%	C5	17.74%	D5	28.80%

Figure 2.5: Interest rates given rating

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 substantial 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.

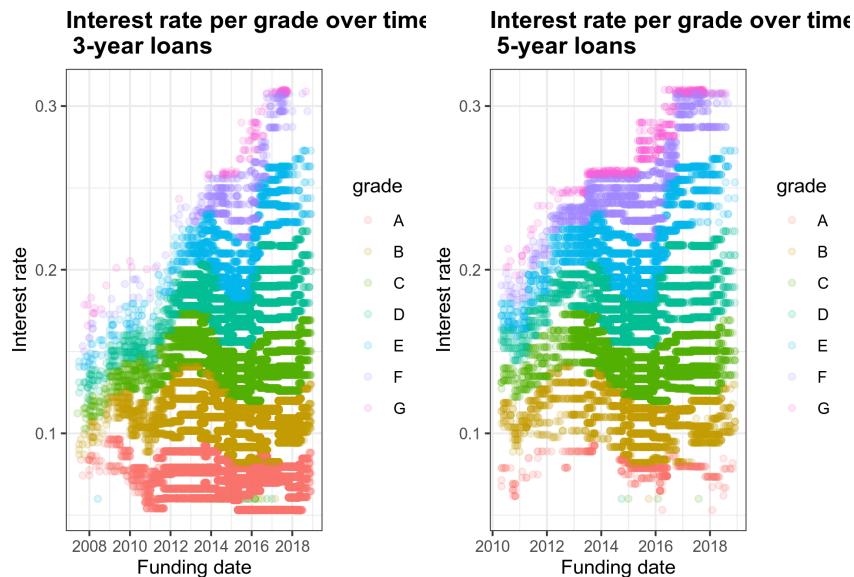


Figure 2.6: Interest rate per grade over time

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).

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

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

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.

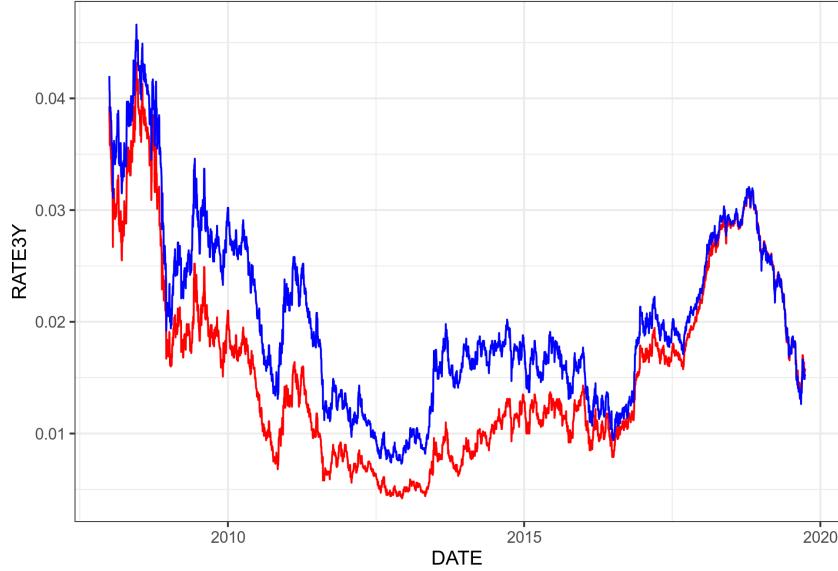


Figure 2.7: Historical Swap Rates

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.

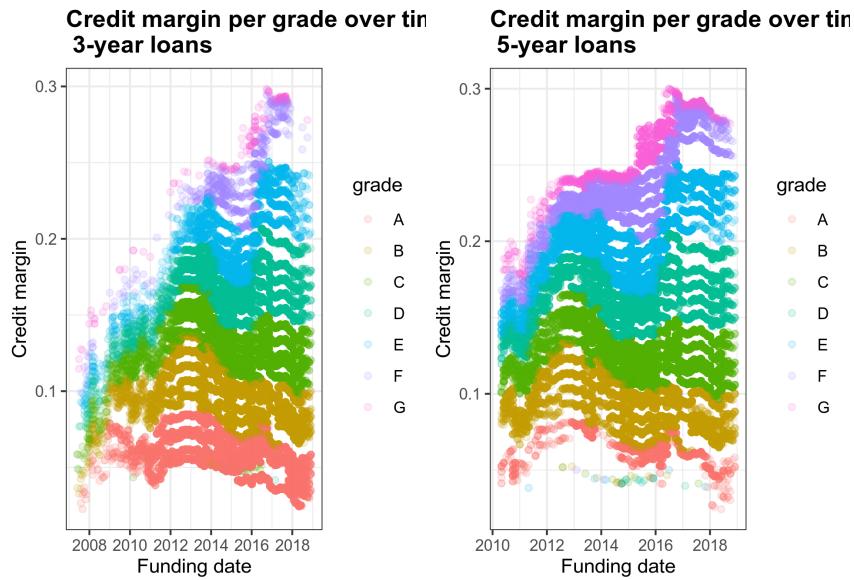


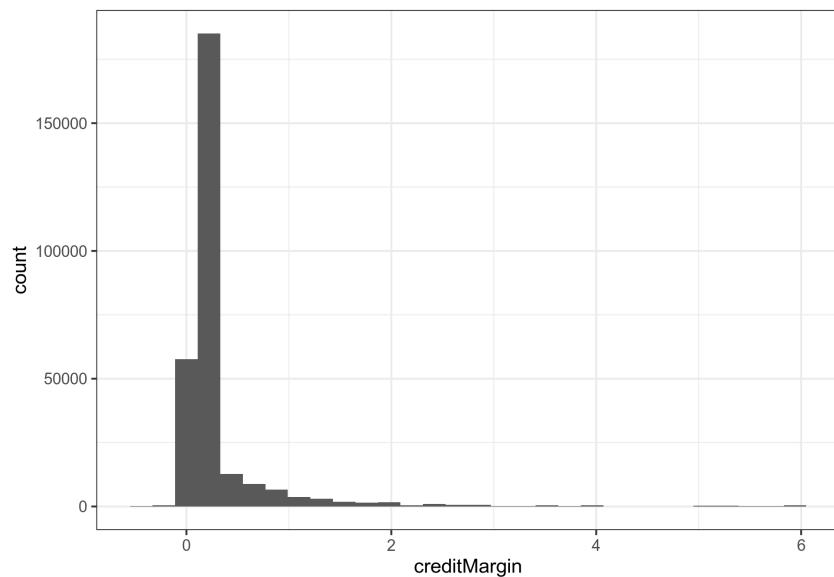
Figure 2.8: Credit margins per grade over time

2.3.4 Purpose

When applying, a potential borrower must state the purpose of the loan. As shown in table 2.2, the main purpose is the consolidation of existing debts.

Table 2.2: Number of loans per purpose

Borrowing purpose	Count
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



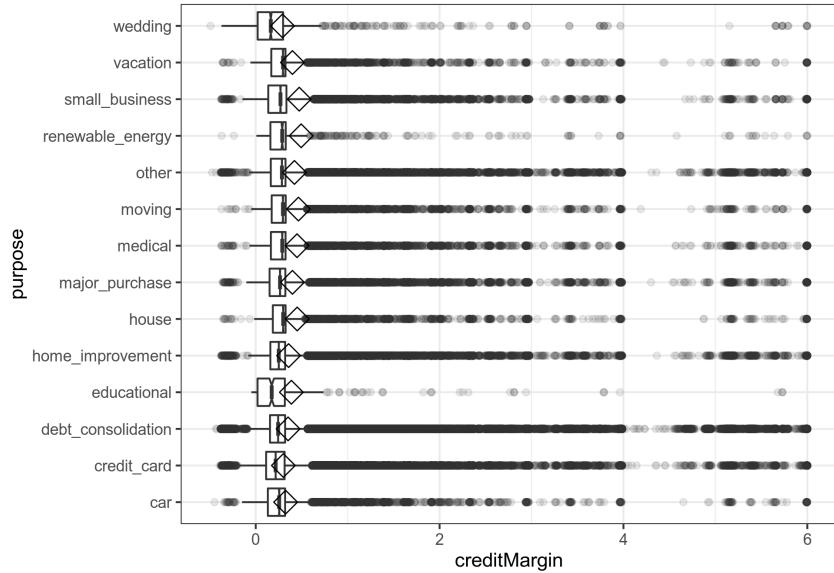


Figure 2.10: Boxplots of credit margins per purpose

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) ^ term), 2),
5       PMT_delta = abs(installment - PMT)
6     ) %>%
7     select(PMT_delta)
8
9   round(mean(100 * installmentError$PMT_delta), digits = 2)
10 }
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 2.11 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.



Figure 2.11: Average NPV et credit margin (%) depending on sub-rating

2.4.2 Principal losses

2.4.3 Distribution of principal losses by rating

Figure 2.12 shows that for a given grade, the losses are very widely spread. The loans are group by ratings and loans that have been fully repaid are removed.

Setting aside the loans rated “A” or “B”, the distributions seem log-normal. Unsurprisingly, the worse the rating the larger the principal loss.

2.4.4 NPV distribution by rating

Principal loss does not reflect the timing of that loss: a loss now is worse than a loss later. This subsection looks at the NPVs of actual loan cashflow (principal and interest) discounted the risk-free rate.

Figure 2.13 shows that for a given grade, the NPVs are very widely spread. From top to bottom, loans are group 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

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.

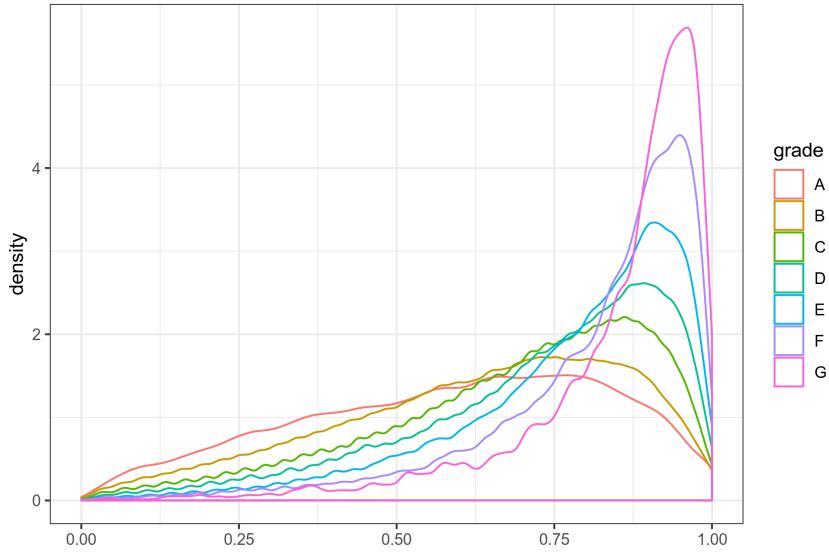


Figure 2.12: Distribution of NPV (%) depending on rating (y-axis square-root scaling)

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.

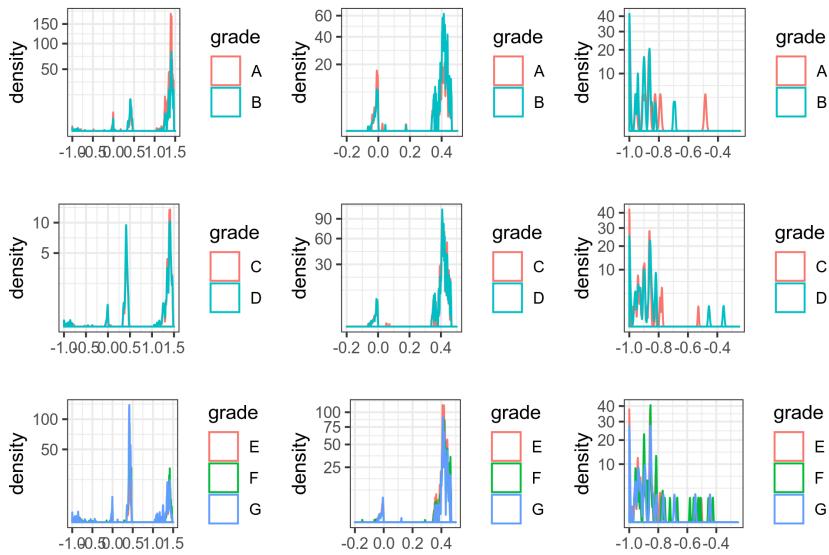


Figure 2.13: Distribution of NPV (%) depending on rating (y-axis square-root scaling)

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

Zooming without scaling the y-axis and grouping all the ratings available for investment on a single plot gives more details.

- Figure 2.14 shows 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);

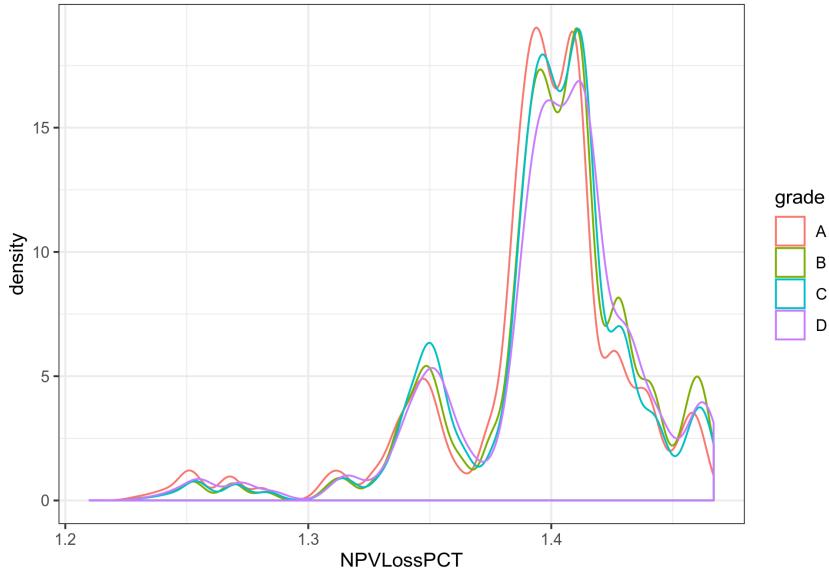


Figure 2.14: NPV % over 120% (no y-axis scaling)

- Figure 2.15 and figure 2.16 show a second and third mode around 41% and -1%;

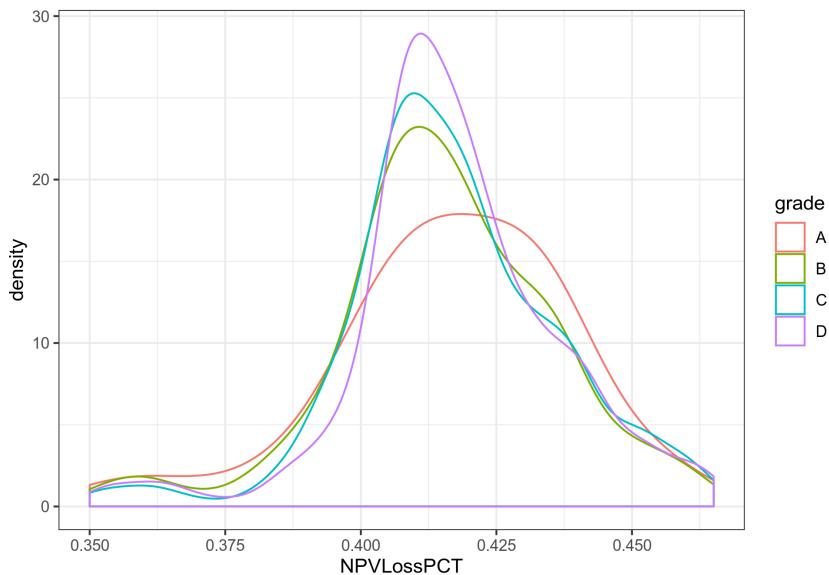


Figure 2.15: NPV % around 41% (no y-axis scaling)

- Finally, figure 2.17 one last very diffuse mode around -100%.

The overall trend is what we should expect. What is surprising is the existence of (1) very clearly defined modes which (2) are common to all types of borrowers. They roughly look log-normal, apart from the mode around 41% which look Gaussian.

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?*

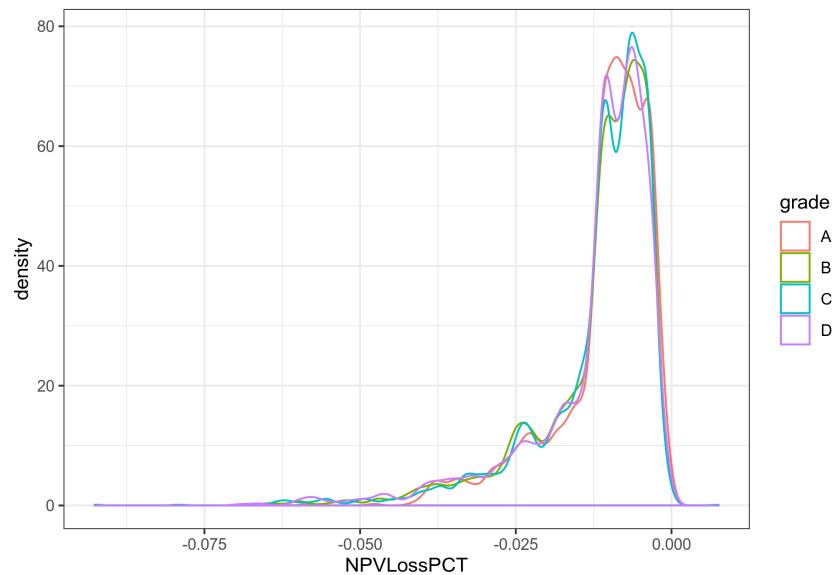


Figure 2.16: NPV % around -1% (no y-axis scaling)

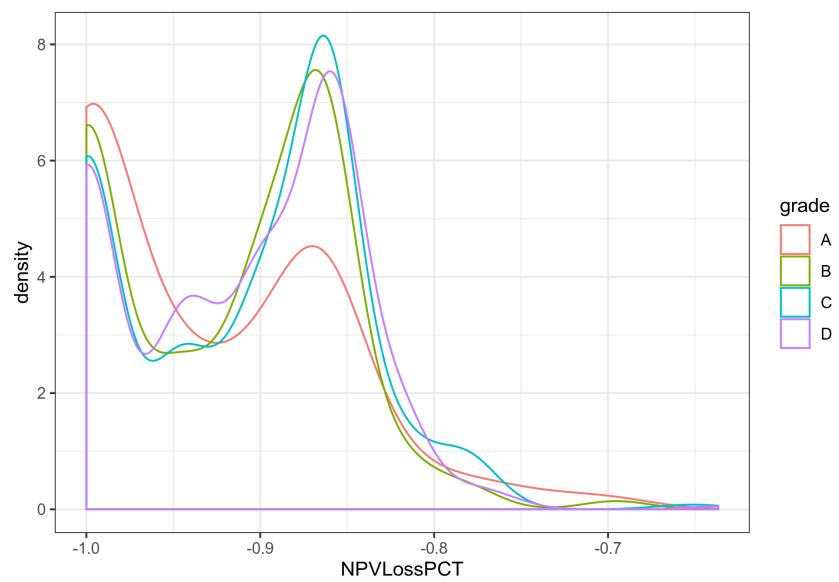


Figure 2.17: NPV % for close to total loss % (no y-axis scaling)

Table 2.3: 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 2.3.

When grouped by grade (Figure 2.18), 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.

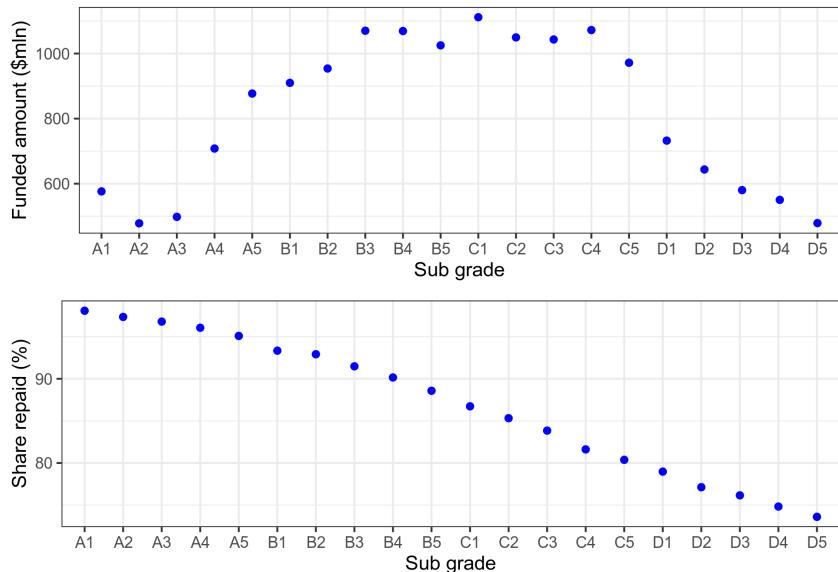


Figure 2.18: Funding and Write-offs by Sub-grades

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

Logistic Regression Model and Credit Scorecard

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.

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?

3.2 Logistic Regression

Logistic models (also called **logit model**) are used to model binary events. Examples would be passing or failing an exam, a newborn being a boy or a girl, a voter choosing a particular political party, or – relevant to us – a borrower defaulting or not on a loan. If the binary variable is modeled as a 0/1 outcome, the model will yield a value between 0 and 1 which can be used as a probability.

We are interested in using a number of variables (being continuous and/or categorical) to model the binary response. A natural model is a linear combination of the variables. Since the predicted value would be continuous and not be bounded by 0/1, the outcome is transformed. A commonly used transformation of the logodds (logarithm of the odds given a particular probability) $\log\left(\frac{p}{1-p}\right)$. This expression has a few advantages: it converts any value (between $-\infty$ and $+\infty$ produced by the linear regression), and it is symmetrical around $x = 0$ and $y = 1/2$. That is, using the odds instead of the probability avoids infinity; it behaves identically when p approaches 0 or 1 (this would not be the case using p). The reciprocal of the logodds is $p = \frac{1}{1+e^{-x}}$.

For a number of X_i variables, the model to fit is then:

$$p = \frac{1}{1 + e^{-\sum_i \alpha_i X_i}}$$

The commonly used method to evaluate the creditworthiness of a borrower is to create scorecards whereby particular characteristics are segmented into intervals and attributed discrete scores. In plain English, a continuous variable (say the age of the applicant) is segmented into intervals (e.g. 0-18 year-olds, 18-26 year-olds, ...), and then given a score. Those different segments become categorical variables. The task of the model is to:

- identify the best way to segment a continuous variable to maximise the information value of the different segments (called bins): intuitively empty or quasi-empty bins (either no or few applicants in the bin) are not informative; bins for which the response is completely random are not predictive, whereas a bin where the response always has the same response is informative (i.e. anybody with income of 0 and 10 dollars per year will default, anybody with a salary of \$1 million per month will repay).
- use a generalised linear model using the new categorical variables;
- transforms the linear coefficients estimated for each category into numerical scores.

WARNING: To run this model, you will need at least 32GB of memory (real plus virtual/swap space)

3.2.1 Data preparation

WARNING: Execute `CleanLoad.Rmd` first

```
1 ## #####  
2 ##  
3 ## Prepare a dataset with ONLY the predictors NOT removing NA's  
4 ##  
5 loansPredictors <-  
6   loansWorkingSet %>%  
7  
8   # Keep the chosen predictors  
9   # Use tidyselect::one_of() to avoid errors if column does not exist
```

```

10  select(tidyselect::one_of(modelVarsIn)) %>%
11
12  ##
13  ## Dates to numeric, in 'decimal' years since 2000
14  ##
15  mutate_at(c("issue_d", "earliest_cr_line"), function(d) {
16    return(year(d) - 2000 + (month(d) - 1) / 12)
17  }) %>%
18
19  ## Add polynomials of the dates to model the time-trend shape
20  mutate(
21    issue_d2 = issue_d^2,
22    issue_d3 = issue_d^3,
23    earliest_cr_line2 = earliest_cr_line^2,
24    earliest_cr_line3 = earliest_cr_line^3
25  ) %>%
26
27  ## Create a logical flag TRUE for non-defaulted (good) loans
28  mutate(isGoodLoan = (principal_loss_pct < 0.001)) %>%
29
30  select(-tidyselect::one_of(modelVarsOut))
31
32
33
34  #####
35  ##
36  ## Create training / test sets 80%/20%
37  ##
38 proportionTraining <- 0.8
39 set.seed(42)
40
41 nSamples <- nrow(loansPredictors)
42
43 sampleTraining <- sample(1:nSamples, floor(nSamples * proportionTraining), replace = FALSE)
44 loansTraining <- loansPredictors %>% slice(sampleTraining)
45 loansTest <- loansPredictors %>% slice(-sampleTraining)
46
47 # Subsets of the training set
48 set.seed(42)
49 nSamplesTraining <- nrow(loansTraining)
50
51 # 1%
52 sample01 <- sample(1:nSamplesTraining, floor(nSamplesTraining * 0.01), replace = FALSE)
53 loans01 <- loansTraining %>% slice(sample01)
54
55 # 20%
56 sample20 <- sample(1:nSamplesTraining, floor(nSamplesTraining * 0.20), replace = FALSE)
57 loans20 <- loansTraining %>% slice(sample20)
58
59
60 ##          used     (Mb) gc trigger     (Mb)    max used     (Mb)
61 ## Ncells   2721369  145.4    4171081   222.8    3814313  203.8
62 ## Vcells  393877605 3005.1  1098756461  8382.9  1098649932 8382.1

```

3.3 Binning and Weight of Evidence

This subsection owes a debt to the source code of the `smbinning` package from which we reimplemented some aspects using the `tidyverse` *style* (the original source code uses SQL

statements to access dataframes), and the documentation vignette of the `Information` package ¹. Our code is provided and imported as a package located on github ².

To identify the various bins (and number and location)

Let's say that we have a binary dependent variable Y and a set of predictive variables X₁,X_p. As mentioned above, Y can capture a wide range of outcomes, such as defaulting on a loan, clicking on an ad, or terminating a subscription.

WOE and IV play two distinct roles when analyzing data:

WOE describes the relationship between a predictive variable and a binary target variable.
IV measures the strength of that relationship.

```
1 library(binner)
```

3.3.1 Loop through all variables

```
1 loansBinning <- loansTraining %>%
2   mutate(
3     home_ownership = as_factor(home_ownership),
4     emp_length = as_factor(emp_length),
5     grade = as_factor(grade)
6   )
7
8 informationValues <- tibble(variable = names(loansBinning), IV = -1.0)
9
10 listBins <-
11   tibble(
12     variable = "",
13     type = "",
14     IV = 0.0,
15     WoE = list(),
16     .rows = 0
17   )
# About 500 sec wall-time
2 startTime <- proc.time()
3 for (n in informationValues$variable) {
4   # cat("Variable: ", n)
5
6   # We don't test the response with itself
7   if (n %in% c("isGoodLoan")) {
8     # cat("\n")
9
10 } else {
11   if (class(loansBinning[[1, n]]) == "factor") {
12     # cat(" is a factor, ")
13     result <- WoETableCategorical(
14       df = loansBinning,
15       x = n,
16       y = "isGoodLoan",
17       maxCategories = 100)
18
19 } else {
20   # cat(" is numeric, ")
```

¹<https://cran.r-project.org/web/packages/Information/vignettes/Information-vignette.html>

²<https://github.com/Emmanuel-R8/SMBinning>

```

21     result <- WoETableContinuous(df = loansBinning,
22                               x = n,
23                               y = "isGoodLoan",
24                               p = 0.05)
25   }
26
27   tryCatch({
28     if (is.na(result)) {
29       # cat("Variable skipped.\n")
30       add_row(
31         listBins,
32         variable = n,
33         type = NA,
34         IV = NA
35       )
36     } else {
37       # cat("    IV : ", result$IV, "\n")
38       listBins <- listBins %>%
39         add_row(
40           variable = n,
41           type = result$type,
42           IV = result$IV,
43           WoE = list(result$table)
44         )
45
46       informationValues[n] <- result$IV
47
48     }
49   },
50   finally = {})
```

```

1 ## Number of missing/NA values: 0
2 ## Number of missing/NA values: 0
3 ## Number of missing/NA values: 0
4 ## Number of missing/NA values: 0
5 ## Number of missing/NA values: 0
6 ## Number of missing/NA values: 247
7 ## Number of missing/NA values: 23
8 ## Number of missing/NA values: 23
9 ## Number of missing/NA values: 24
10 ## Number of missing/NA values: 527124
11 ## Number of missing/NA values: 866836
12 ## Number of missing/NA values: 23
13 ## Number of missing/NA values: 23
14 ## Number of missing/NA values: 0
15 ## Number of missing/NA values: 687
16 ## Number of missing/NA values: 23
17 ## Number of missing/NA values: 645401
18 ## Number of missing/NA values: 645401
19 ## Number of missing/NA values: 645401
20 ## Number of missing/NA values: 645401
21 ## Number of missing/NA values: 56032
22 ## Number of missing/NA values: 51160
23 ## Number of missing/NA values: 51160
24 ## Number of missing/NA values: 85391
25 ## Number of missing/NA values: 56016
```

```

26 ## Number of missing/NA values: 56016
27 ## Number of missing/NA values: 56015
28 ## Number of missing/NA values: 39839
29 ## Number of missing/NA values: 49826
30 ## Number of missing/NA values: 797767
31 ## Number of missing/NA values: 137406
32 ## Number of missing/NA values: 696281
33 ## Number of missing/NA values: 56015
34 ## Number of missing/NA values: 56015
35 ## Number of missing/NA values: 56015
36 ## Number of missing/NA values: 46723
37 ## Number of missing/NA values: 56015
38 ## Number of missing/NA values: 56015
39 ## Number of missing/NA values: 56015
40 ## Number of missing/NA values: 56015
41 ## Number of missing/NA values: 56015
42 ## Number of missing/NA values: 46723
43 ## Number of missing/NA values: 56015
44 ## Number of missing/NA values: 56015
45 ## Number of missing/NA values: 56139
46 ## Number of missing/NA values: 50862
47 ## Number of missing/NA values: 1114
48 ## Number of missing/NA values: 56015
49 ## Number of missing/NA values: 39839
50 ## Number of missing/NA values: 39839
51 ## Number of missing/NA values: 0
52 ## Number of missing/NA values: 0
53 ## Number of missing/NA values: 23
54 ## Number of missing/NA values: 23

1 proc.time() - startTime

1 ##       user   system elapsed
2 ## 771.141   5.720 777.852

1 saveRDS(informationValues, "datasets/informationValues100.rds")
2 saveRDS(listBins, "datasets/listBins100.rds")

```

3.3.2 Select relevant variables

It is considered that information value below 2% is useless. The 15 best variables (in terms of Information Value) are:

```

1 bestBins <- listBins %>% filter(IV >= 0.02)
2
3 bestBins %>%
4   select(variable, IV) %>%
5   mutate(IV = round(IV, digits = 5)) %>%
6   arrange(desc(IV)) %>%
7   slice(1:15)

```

variable	IV
sub_grade	0.49392
int_rate	0.46527
grade	0.46016
term	0.17329
dti	0.07445
verification_status	0.05681
avg_cur_bal	0.05430
tot_hi_cred_lim	0.05003
num_tl_op_past_12m	0.04852
mort_acc	0.04382
total_bc_limit	0.04118
issue_d2	0.04087
issue_d3	0.04087
issue_d	0.04083
loan_amnt	0.03616

However, we notice that this list includes variables that would not be available at the time the credit scoring is performed.

```

1 bestBins <- bestBins %>%
2   filter(!(variable %in% c(
3     "loanID", "term", "int_rate", "creditMargin", "loan_status",
4     "grade", "sub_grade", "grade_num", "sub_grade_num",
5     "emp_length", "home_ownership", "monthDefault",
6     "principal_loss_pct", "creditMargin", "monthDefault",
7     "isGoodLoan")))
8
9 saveRDS(bestBins, "datasets/bestBins100.rds")

```

The 15 best variables that will retain are in the following table. Interestingly, the square and cubic powers of the issue date are retained.

```

1 bestBins %>%
2   select(variable, IV) %>%
3   mutate(IV = round(IV, digits = 5)) %>%
4   arrange(desc(IV)) %>%
5   slice(1:15)

```

variable	IV
dti	0.07445
verification_status	0.05681
avg_cur_bal	0.05430
tot_hi_cred_lim	0.05003
num_tl_op_past_12m	0.04852
mort_acc	0.04382
total_bc_limit	0.04118
issue_d2	0.04087
issue_d3	0.04087
issue_d	0.04083
loan_amnt	0.03616
mths_since_recent_inq	0.03488
mo_sin_rcnt_tl	0.03359
num_actv_rev_tl	0.03354
open_rv_24m	0.03338

And the 10 worst (but retained) variables are:

```

1 bestBins %>%
2   select(variable, IV) %>%

```

```

3   mutate(IV = round(IV, digits = 5)) %>%
4     arrange(IV) %>%
5     slice(1:15)

```

variable	IV
max_bal_bc	0.02120
revol_util	0.02549
open_rv_12m	0.02650
inq_last_6mths	0.02654
mo_sin_old_rev_tl_op	0.02705
mths_since_recent_bc	0.02727
mo_sin_rcnt_rev_tl_op	0.02959
bc_open_to_buy	0.03048
bc_util	0.03048
num_rev_tl_bal_gt_0	0.03221
percent_bc_gt_75	0.03257
open_rv_24m	0.03338
num_actv_rev_tl	0.03354
mo_sin_rcnt_tl	0.03359
mths_since_recent_inq	0.03488

3.3.3 Create data table with only binary variables (transform every bin to a 0/1 value)

For each variable, create new variables for each bin in the WoE table of that variable. Strictly speaking, this is not necessary for the generalised model algorithms. They are able to model using categories containing the factors. However, as we will see, the model will generate a number of NAs for variable (factors) which are co-linear. (This is the case for any linear model.) This will require removing those individual factors, which we found easier to do when each factor is given an individual variable (column).

```

1 # Variable will contain all the best characteristics. Every continuous variable is reformatted
2 # into factors reflecting the appropriate bins
3
4 allFactorsAsCharacteristics <- loansTraining[, "loanID"]
5 allFactorsAsBins <- loansTraining[, "loanID"]

1 for (index in 1:nrow(bestBins)) {
2   #for (index in 1:27) {
3     name <- bestBins$variable[index][[1]]
4
5     cat("-----",
6       index,
7       "--",
8       name, "\n")
9
10    ltIndex <- which(names(loansTraining) == name)
11
12    characteristic <- categoriseFromWoE(df = loansTraining[, ltIndex],
13                                         varName = name,
14                                         woeTable = bestBins$WoE[[index]][[1]])
15
16    bins <- categoriseFromWoE.Wide(df = loansTraining[, ltIndex],
17                                      varName = name,
18                                      woeTable = bestBins$WoE[[index]][[1]])
19
20  allFactorsAsCharacteristics <-

```

```

21     allFactorsAsCharacteristics %>%
22       cbind(characteristic)
23
24
25   allFactorsAsBins <-
26     allFactorsAsBins %>%
27       cbind(bins)
28
29 }

```

```

1 ## ----- 1 -- loan_amnt
2 ## ----- 2 -- verification_status
3 ## ----- 3 -- issue_d
4 ## ----- 4 -- dti
5 ## ----- 5 -- inq_last_6mths
6 ## ----- 6 -- revol_util
7 ## ----- 7 -- open_rv_12m
8 ## ----- 8 -- open_rv_24m
9 ## ----- 9 -- max_bal_bc
10 ## ----- 10 -- avg_cur_bal
11 ## ----- 11 -- bc_open_to_buy
12 ## ----- 12 -- bc_util
13 ## ----- 13 -- mo_sin_old_rev_tl_op
14 ## ----- 14 -- mo_sin_rcnt_rev_tl_op
15 ## ----- 15 -- mo_sin_rcnt_tl
16 ## ----- 16 -- mort_acc
17 ## ----- 17 -- mths_since_recent_bc
18 ## ----- 18 -- mths_since_recent_inq
19 ## ----- 19 -- num_actv_rev_tl
20 ## ----- 20 -- num_rev_tl_bal_gt_0
21 ## ----- 21 -- num_tl_op_past_12m
22 ## ----- 22 -- percent_bc_gt_75
23 ## ----- 23 -- tot_hi_cred_lim
24 ## ----- 24 -- total_bc_limit
25 ## ----- 25 -- issue_d2
26 ## ----- 26 -- issue_d3

```

3.3.4 Comparison of individual characteristics

- re-train secures the lowest Akaike criterion and will be chosen.

- On this test, the best variables are: SORT BY IV. Then WoE for the first few ones.

```

1 loansBinning <- loans20 %>%
2   mutate(
3     home_ownership = as_factor(home_ownership),
4     emp_length = as_factor(emp_length)
5   )

```

The following plots the 10 top variables weight of evidence plots. They show for those variables the weight of evidence of each individual factor. A positive WoE shows that the factor is positive to explain a positive outcome (that is that a loan is good).

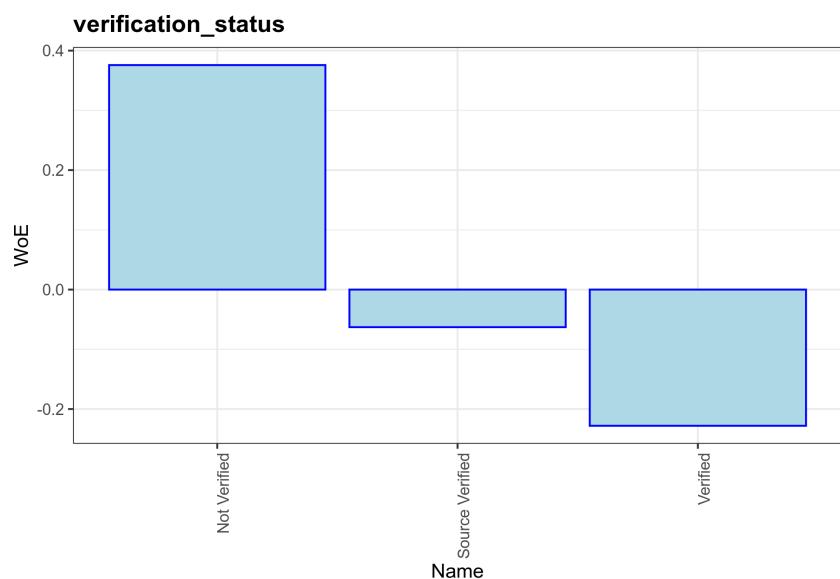
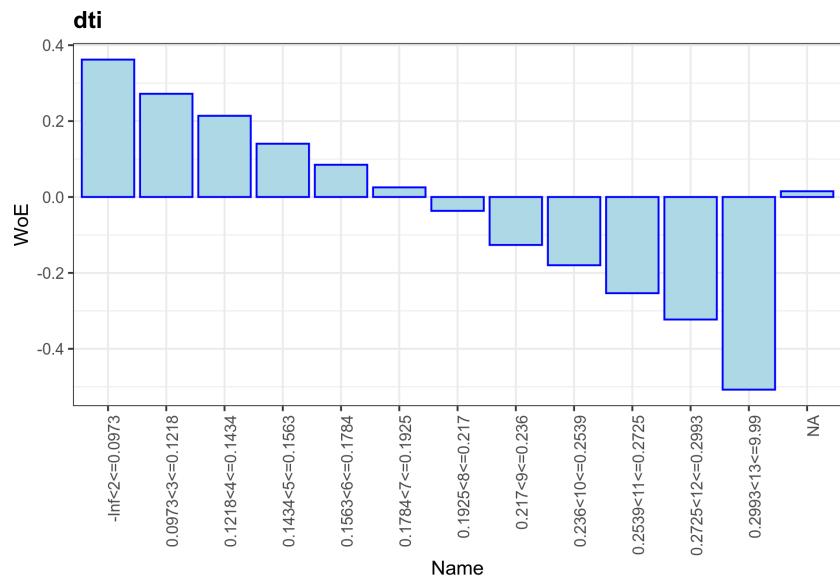
```

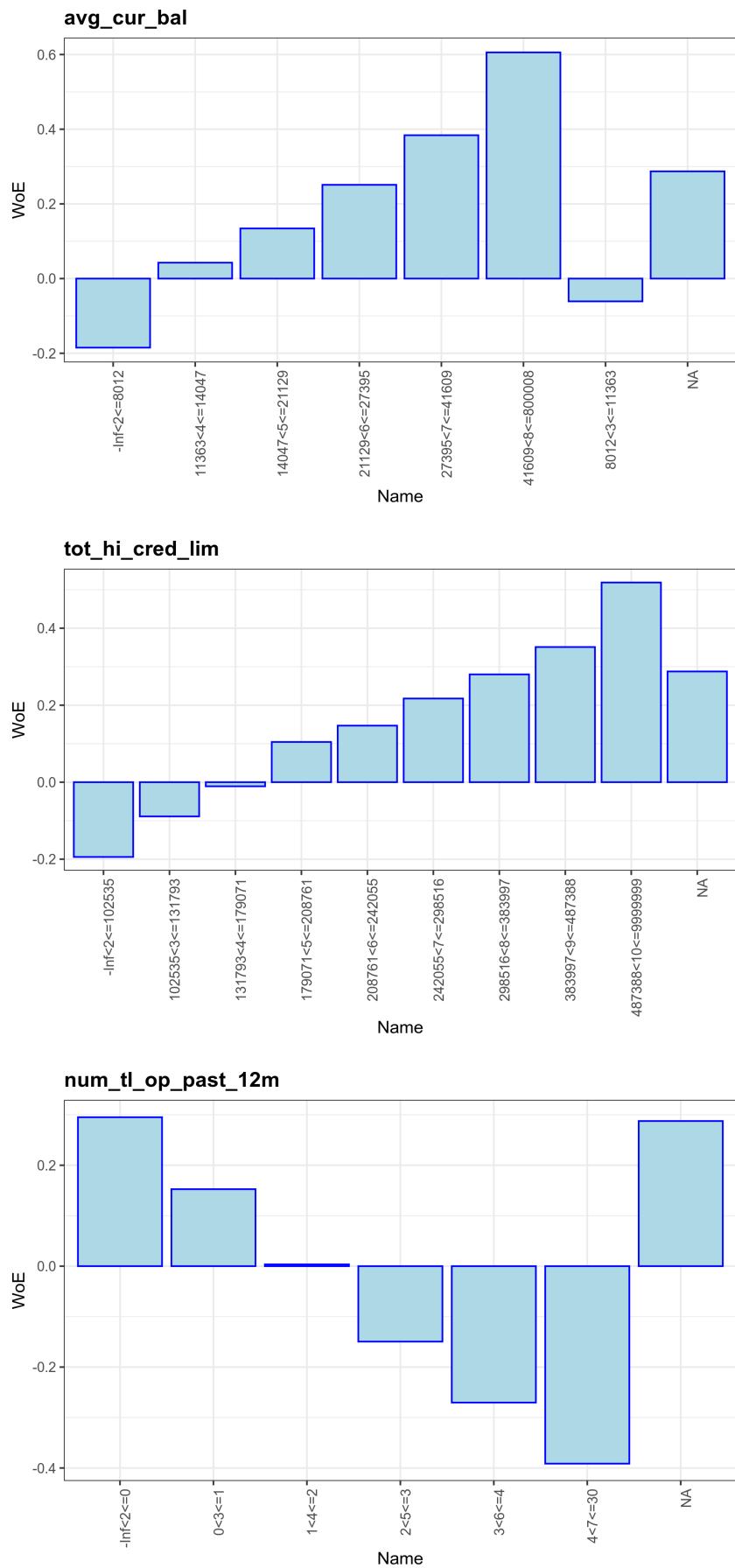
1 best10 <- bestBins %>% arrange(desc(IV)) %>% slice(1:10)
2
3 plotBinWoE <- function(n = 1) {
4   vName <- best10[n,]$variable
5
6   if (best10[n,]$type == "numeric") {

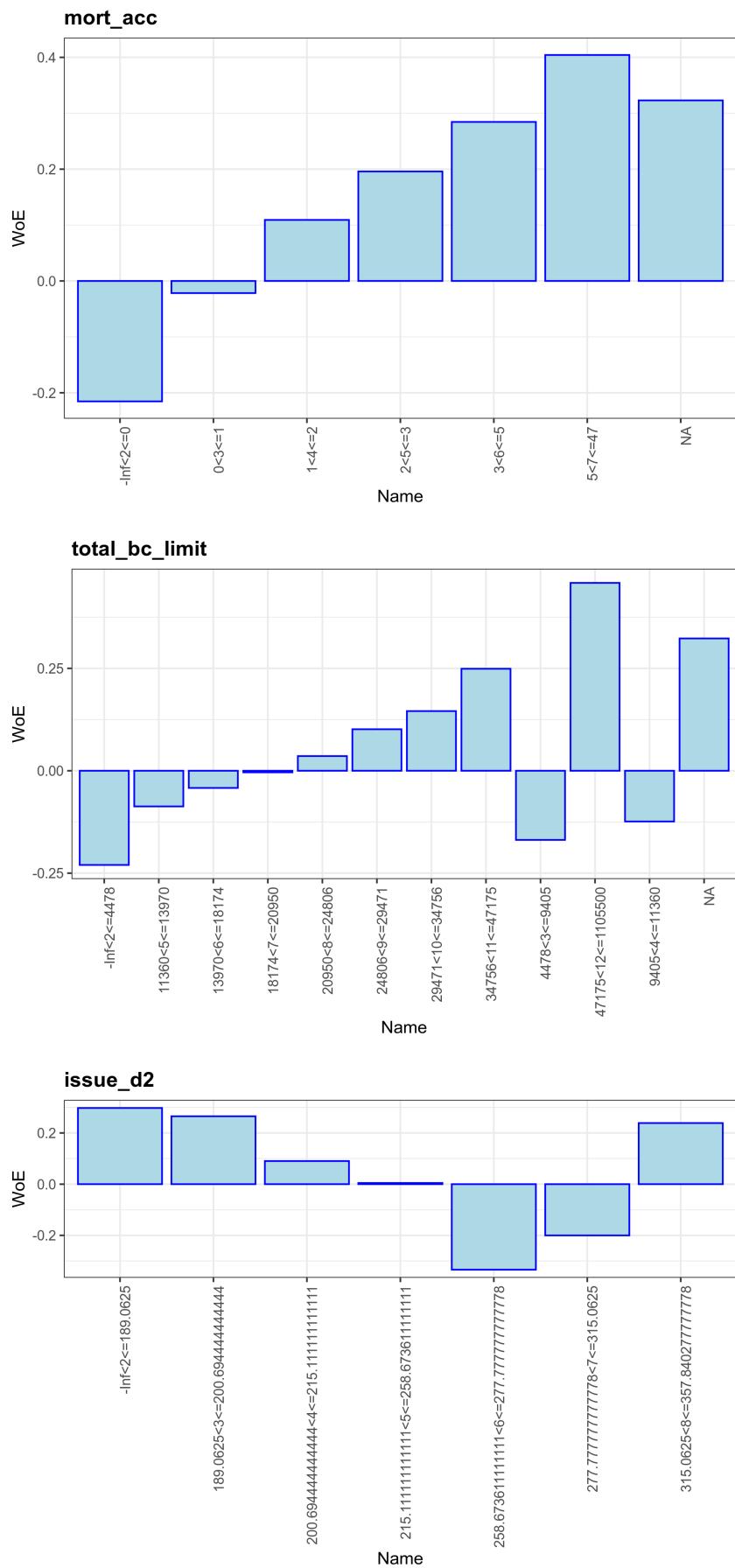
```

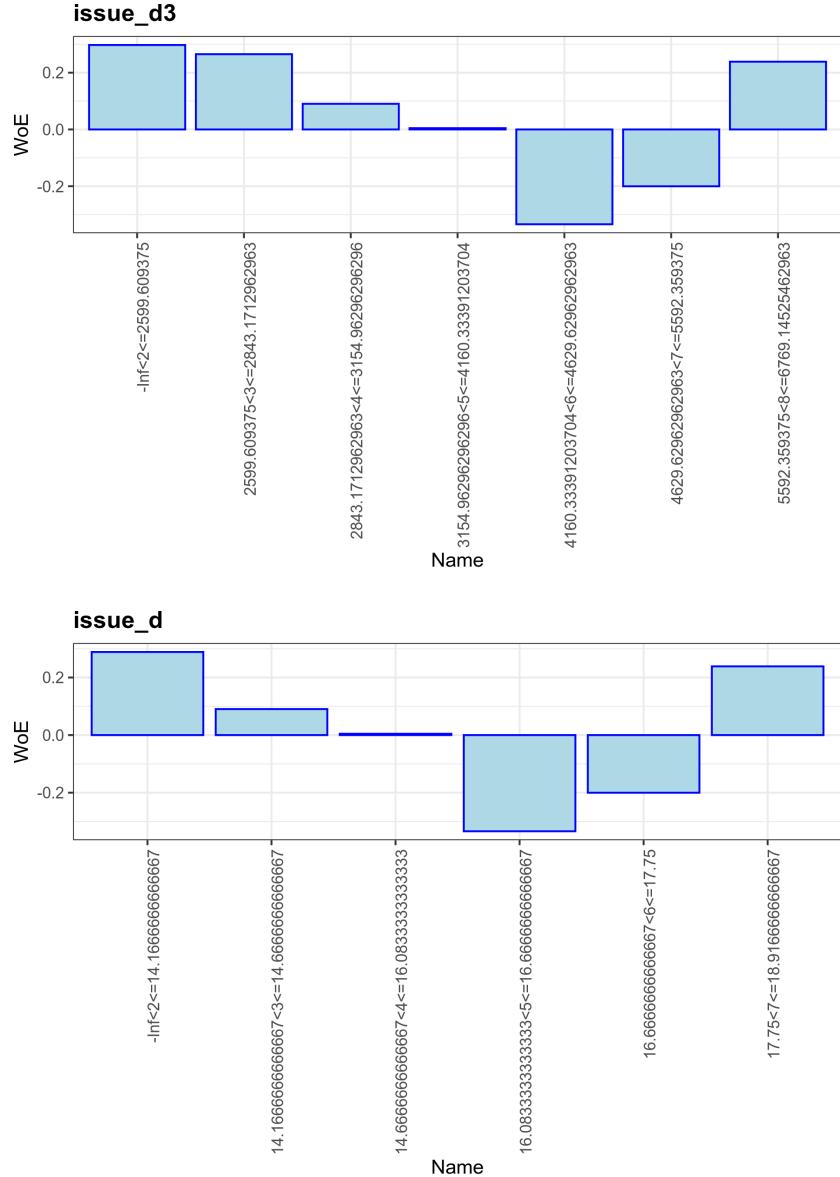
```

7   best10[n,]$WoE[[1]] %>%
8     arrange(WoE) %>%
9     ggplot(aes(Name, WoE)) +
10    geom_col(col = "blue", fill = "lightblue") +
11    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
12    ggtitle(vName)
13 } else {
14   best10[n,]$WoE[[1]] %>%
15     arrange(WoE) %>%
16     ggplot(aes(Name, WoE)) +
17     geom_col(col = "blue", fill = "lightblue") +
18     theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
19     ggtitle(vName)
20 }
21
22 }
23 }
```









3.4 Logistic Regression

3.4.1 Logistic regression using the SpeedGLM package

NOTE: `speedglm` has a `select()` function which would shadow `dplyr::select()`, so it is only used fully qualified to avoid any collision.

```

1 if (!exists("allFactorsAsBins")) {
2   allFactorsAsBins <- readRDS("datasets/allBins100.rds")
3 }
4
5 loanSampleBins <-
6   allFactorsAsBins %>%
7   select(-loanID) %>%
8   cbind(loansTraining$isGoodLoan) %>%
9   rename(isGoodLoan = "loansTraining$isGoodLoan") %>%
10  mutate(isGoodLoan = if_else(isGoodLoan, 1, 0)) %>%
11  as.data.frame()
12

```

```

13 rm(allFactorsAsBins)
14
15
16
17
18 if (!exists("allFactorsAsCharacteristics")) {
19   allFactorsAsCharacteristics <-
20     readRDS("datasets/allCharacteristics100.rds")
21 }
22
23 loanSampleCharacteristics <-
24   allFactorsAsCharacteristics %>%
25   select(-loanID) %>%
26   cbind(loansTraining$isGoodLoan) %>%
27   rename(isGoodLoan = "loansTraining$isGoodLoan") %>%
28   mutate(isGoodLoan = if_else(isGoodLoan, 1, 0)) %>%
29   as.data.frame()
30
31 rm(allFactorsAsCharacteristics)

1 ##                  used      (Mb) gc trigger      (Mb)    max used      (Mb)
2 ## Ncells    2775809  148.3    4554342  243.3    4554342  243.3
3 ## Vcells   394543674 3010.2  1098756461 8382.9  1098696277 8382.4

1 ## [1] 222
1 ## [1] 27

```

3.4.2 Remove identical bins

```

1 namesCharacteristics <- names(loanSampleCharacteristics)
2 namesBins <- names(loanSampleBins)
3
4
5 # Starting list of names
6 # > The following two variables are only useful if all the bins have been given their own column (if
7 # > characteristic has 7 factors, that creates 7 bins and column).
8
9
10 # Running GLM would give a number of NAs if some columns are linearly dependant (in our case
11 # actually equal). A "trick" to identify them is to run a hash digest on each column and spot
12 # identical hashes. The original variables were not identical, but when split into categories, they
13 # can become identical. For example, if no income is provided (income=NA category), no
14 # debt-to-income ratio can be calculated (dti=NA).
15 duplicateNames <- loanSampleBins[duplicated(lapply(loanSampleBins, digest::digest))] %>% names()
16 duplicateNames

1 ## [1] "(open_rv_24m) open_rv_24m=NA"                                "(max_bal_bc) max_bal_bc=NA"
2 ## [3] "(bc_util) -Inf<bc_util<=31.5"                                "(bc_util) 31.5<bc_util<=38.8"
3 ## [5] "(bc_util) 38.8<bc_util<=46.1"                                "(bc_util) 46.1<bc_util<=53.9"
4 ## [7] "(bc_util) 53.9<bc_util<=72.5"                                "(bc_util) 72.5<bc_util<=87.6"
5 ## [9] "(bc_util) 87.6<bc_util<=91.9"                                "(bc_util) 91.9<bc_util<=97.9"
6 ## [11] "(bc_util) 97.9<bc_util<=339.6"                               "(bc_util) bc_util=NA"
7 ## [13] "(mo_sin_rcnt_rev_tl_op) mo_sin_rcnt_rev_tl_op=NA"      "(num_actv_rev_tl) num_actv_rev_tl=NA"
8 ## [15] "(num_rev_tl_bal_gt_0) num_rev_tl_bal_gt_0=NA"          "(num_tl_op_past_12m) num_tl_op_past_12m=NA"
9 ## [17] "(tot_hi_cred_lim) tot_hi_cred_lim=NA"                  "(total_bc_limit) total_bc_limit=NA"
10 ## [19] "(issue_d2) 200.694<issue_d2<=215.111"                "(issue_d2) 215.111<issue_d2<=258.674"
11 ## [21] "(issue_d2) 258.674<issue_d2<=277.778"                "(issue_d2) 277.778<issue_d2<=315.062"

```

```

12 ## [23] "(issue_d2) 315.062<issue_d2<=357.84"          "(issue_d3) -Inf<issue_d3<=2599.609"
13 ## [25] "(issue_d3) 2599.609<issue_d3<=2843.171"      "(issue_d3) 2843.171<issue_d3<=3154.963"
14 ## [27] "(issue_d3) 3154.963<issue_d3<=4160.334"       "(issue_d3) 4160.334<issue_d3<=4629.63"
15 ## [29] "(issue_d3) 4629.63<issue_d3<=5592.359"        "(issue_d3) 5592.359<issue_d3<=6769.145"

1 # Remove any categories uniformly constant (only 0)
2 #
3 # This is important when working on a small extract of the dataset for variables that are seldom
4 # used (e.g. customers from certain states).
5 zeroColumnsNames <- loanSampleBins[, colSums(loanSampleBins) == 0] %>% names()

```

3.4.3 First training on variables previously selected on their Information Value

We use `speedglm` being quick. Note that alternatives were also tried: `glm` crashed on even small extracts of the dataset. `glmnet` returns errors that were not understandable or documented on the internet.

3.4.3.1 GLM on characteristics

```

1 ##           used   (Mb) gc trigger   (Mb) max used   (Mb)
2 ## Ncells    2729716 145.8   8544085  456.4   8544085  456.4
3 ## Vcells   416513987 3177.8 1325866348 10115.6 1325400563 10112.1

1 # About 700 sec wall-time to complete training dataset
2 # The dataset is the sample less duplicate, less zero-ed columns
3 {
4   doMC::registerDoMC(cores = 1) # Too many processes push over 32GB
5   startTime <- proc.time()

6
7   loansData <- loanSampleCharacteristics %>%
8     select(-one_of( c(duplicateNames,
9                   zeroColumnsNames)))
10
11 SGLM_C_train <- speedglm::speedglm(isGoodLoan ~ .,
12                                     data = loansData,
13                                     family = binomial())
14
15 doMC::registerDoMC(cores = NULL)
16 cat(proc.time() - startTime, "\n")
17 }

1 ## 148.955 3.265 152.248 0 0
1 #summary(SGLM_C_train)
2
3 # There are a number of NA estimates that need to be removed by removing individual bins
4 length( which(is.na(summary(SGLM_C_train)$coefficient$Estimate)) )

1 ## [1] 28
1 saveRDS(SGLM_C_train, "datasets/SGLM_C_train.rds")

```

3.4.3.2 GLM on bins

```

1 ##           used   (Mb) gc trigger   (Mb) max used   (Mb)
2 ## Ncells    2729879 145.8   8544085  456.4   8544085  456.4
3 ## Vcells   416514817 3177.8 1325866348 10115.6 1325400563 10112.1

```

```

1 # About 700 sec wall-time to complete training dataset
2 # The dataset is the sample less duplicate, less zero-ed columns
3 {
4   doMC::registerDoMC(cores = 1) # Too many processes push over 32GB
5   startTime <- proc.time()
6
7   loansData <- loanSampleBins
8
9   SGLM_B_train <- speedglm::speedglm(isGoodLoan ~ .,
10                                         data = loansData,
11                                         family = binomial())
12
13  doMC::registerDoMC(cores = NULL)
14  cat(proc.time() - startTime, "\n")
15 }

## 172.621 9.823 208.567 0 0

1 saveRDS(SGLM_B_train, "datasets/SGLM_B_train.rds")
2
3 # There are a number of NA estimates that need to be removed by removing individual bins

1 speedglm:::summary.speedglm(SGLM_B_train)$coefficients %>%
2   as.data.frame() %>%
3   rownames_to_column(var = "binName") %>%
4   filter(is.na(Estimate)) %>%
5   slice(1:10)

```

binName	Estimate	Std. Error	z value	Pr(> z)
'(loan_amnt) 28000<loan_amnt<=40000'	NA	NA	NA	NA
'(verification_status) verification_status=Verified'	NA	NA	NA	NA
'(issue_d) 17.75<issue_d<=18.917'	NA	NA	NA	NA
'(dti) dti=NA'	NA	NA	NA	NA
'(inq_last_6mths) inq_last_6mths=NA'	NA	NA	NA	NA
'(revol_util) revol_util=NA'	NA	NA	NA	NA
'(open_rv_12m) open_rv_12m=NA'	NA	NA	NA	NA
'(open_rv_24m) 5<open_rv_24m<=53'	NA	NA	NA	NA
'(open_rv_24m) open_rv_24m=NA'	NA	NA	NA	NA
'(max_bal_bc) 7905<max_bal_bc<=776843'	NA	NA	NA	NA

```

1 speedglm:::summary.speedglm(SGLM_B_train)$coefficients %>%
2   as.data.frame() %>%
3   rownames_to_column(var = "binName") %>%
4   rename(p = "Pr(>|z|)",
5         z = "z value") %>%
6   arrange(desc(z)) %>%
7   slice(1:10)

```

binName	Estimate	Std. Error	z	p
'(loan_amnt) 3500<loan_amnt<=9000'	0.9848386	0.0108316	90.9228	0.00e+00
'(loan_amnt) -Inf<loan_amnt<=3500'	1.1617352	0.0152225	76.3171	0.00e+00
'(loan_amnt) 9000<loan_amnt<=10000'	0.7380869	0.0124309	59.3751	0.00e+00
'(loan_amnt) 10000<loan_amnt<=12000'	0.6002188	0.0121587	49.3652	0.00e+00
'(verification_status) verification_status=Not Verified'	0.2840857	0.0070796	40.1275	0.00e+00
'(loan_amnt) 14975<loan_amnt<=15000'	0.5501235	0.0142603	38.5774	0.00e+00
'(loan_amnt) 12000<loan_amnt<=14975'	0.5037774	0.0131783	38.2277	0.00e+00
'(num_tl_op_past_12m) 0<num_tl_op_past_12m<=1'	0.3467503	0.0121796	28.4698	2.77e-178
'(loan_amnt) 15000<loan_amnt<=17000'	0.3762410	0.0135998	27.6652	1.83e-168
'(mort_acc) 5<mort_acc<=47'	1.0623545	0.0423491	25.0856	7.14e-139

```

1 # Does the model throw NAs? They are produced not only for identical variables, but also co-linear
2 # combinations. Let us select those variables:
3
4 NAsFirstTraining <-
5
6 # Take the results (just the estimated coefficients) from the model
7 tibble(
8   name = rownames(summary(SGLM_B_train)$coefficient),
9   estimate = summary(SGLM_B_train)$coefficient$Estimate
10 ) %>%
11
12 # Reformat the names to be the same as in the bins
13 mutate(name = stringr::str_remove_all(name, "\`")) %>%
14
15 # Make sure that response is not deleted by mistake and list all NAs
16 filter(name != "isGoodLoan" & is.na(estimate))
17
18 NAsFirstTraining <- NAsFirstTraining$name

```

3.4.4 Second training

```

1 ##          used    (Mb) gc trigger    (Mb)   max used    (Mb)
2 ## Ncells  2735490  146.1   8544085  456.4   8544085  456.4
3 ## Vcells 416569903 3178.2 1272895694 9711.5 1511294518 11530.3
4
5 # Start the model training again
6 {
7   startTime <- proc.time()
8   doMC::registerDoMC(cores = 2) # Too many processes push over 32GB if more than 2
9
10 loansData <- loanSampleBins %>%
11   select(-one_of( c(duplicateNames,
12                   zeroColumnsNames,
13                   NAsFirstTraining)))
14
15 SGLM_B_retrain <- speedglm::speedglm(isGoodLoan ~ .,
16                                         data = loansData,
17                                         family = binomial())
18
19 doMC::registerDoMC(cores = NULL)
20 cat(proc.time() - startTime)
21 }
22
23 ## 142.721 3.661 151.765 0 0
24
25 saveRDS(SGLM_B_retrain, "datasets/SGLM_B_retrain.rds")
26
27 speedglm:::summary.speedglm(SGLM_B_retrain)$coefficients %>%
28   as.data.frame() %>%
29   rownames_to_column(var = "binName") %>%
30   rename(p = "Pr(>|z|)",
31         z = "z value") %>%
32   arrange(desc(z)) %>%
33   slice(1:10)

```

binName	Estimate	Std. Error	z	p
'(loan_amnt) 3500<loan_amnt<=9000'	0.9848386	0.0108316	90.9228	0.00e+00
'(loan_amnt) -Inf<loan_amnt<=3500'	1.1617352	0.0152225	76.3171	0.00e+00
'(loan_amnt) 9000<loan_amnt<=10000'	0.7380869	0.0124309	59.3751	0.00e+00
'(loan_amnt) 10000<loan_amnt<=12000'	0.6002188	0.0121587	49.3652	0.00e+00
'(verification_status) verification_status=Not Verified'	0.2840857	0.0070796	40.1275	0.00e+00
'(loan_amnt) 14975<loan_amnt<=15000'	0.5501235	0.0142603	38.5774	0.00e+00
'(loan_amnt) 12000<loan_amnt<=14975'	0.5037774	0.0131783	38.2277	0.00e+00
'(num_tl_op_past_12m) 0<num_tl_op_past_12m<=1'	0.3467503	0.0121796	28.4698	2.77e-178
'(loan_amnt) 15000<loan_amnt<=17000'	0.3762410	0.0135998	27.6652	1.83e-168
'(mort_acc) 5<mort_acc<=47'	1.0623545	0.0423491	25.0856	7.14e-139

Just in case, new NAs popped up. Then select variables whose significance value is under 2 sigmas.

```

1 NAsSecondTraining <-
2   tibble(
3     name = rownames(summary(SGLM_B_retrain)$coefficient),
4     estimate = summary(SGLM_B_retrain)$coefficient$Estimate,
5     zValue = abs(summary(SGLM_B_retrain)$coefficient$"z value")
6   ) %>%
7
8   mutate(name = stringr::str_remove_all(name, "\`")) %>%
9
10  filter(name != "isGoodLoan") %>%
11
12  # Remove stray NAs or anything less than 2 sigmas
13  filter(is.na(estimate) | zValue < 2)
14
15
16 NAsSecondTraining <- NAsSecondTraining$name

```

3.4.5 Third training

```

1 ## used (Mb) gc trigger (Mb) max used (Mb)
2 ## Ncells 2777129 148.4 4554342 243.3 4554342 243.3
3 ## Vcells 641189458 4891.9 1098756461 8382.9 1098696277 8382.4

# Start the model training again. About 350 sec.
{
  startTime <- proc.time()
  doMC::registerDoMC(cores = 2)

  loansData <- loanSampleBins %>%
    select(-one_of( c(
      duplicateNames,
      zeroColumnNames,
      NAsFirstTraining,
      NAsSecondTraining
    )))

  SGLM_B_reretrain <- speedglm::speedglm(isGoodLoan ~ .,
                                            data = loansData,
                                            family = binomial())

  doMC::registerDoMC(cores = NULL)
  cat(proc.time() - startTime)
}

## 50.463 1.961 56.132 0 0

```

```

1 saveRDS(SGLM_B_reretrain, "datasets/SGLM_B_reretrain.rds")

1 speedglm:::summary.speedglm(SGLM_B_reretrain)$coefficients %>%
2   as.data.frame() %>%
3   rownames_to_column(var = "binName") %>%
4   rename(p = "Pr(>|z|)",
5         z = "z value") %>%
6   arrange(desc(z)) %>%
7   slice(1:10)



| binName                                                  | Estimate  | Std. Error | z       | p         |
|----------------------------------------------------------|-----------|------------|---------|-----------|
| '(loan_amnt) 3500<loan_amnt<=9000'                       | 0.9383885 | 0.0107265  | 87.4833 | 0.00e+00  |
| '(loan_amnt) -Inf<loan_amnt<=3500'                       | 1.1079606 | 0.0151211  | 73.2723 | 0.00e+00  |
| '(loan_amnt) 9000<loan_amnt<=10000'                      | 0.7026804 | 0.0123379  | 56.9531 | 0.00e+00  |
| '(loan_amnt) 10000<loan_amnt<=12000'                     | 0.5616891 | 0.0120636  | 46.5606 | 0.00e+00  |
| '(verification_status) verification_status=Not Verified' | 0.3225105 | 0.0070171  | 45.9606 | 0.00e+00  |
| '(num_tl_op_past_12m) -Inf<num_tl_op_past_12m<=0'        | 0.5700935 | 0.0133125  | 42.8238 | 0.00e+00  |
| '(num_tl_op_past_12m) 0<num_tl_op_past_12m<=1'           | 0.4623390 | 0.0112317  | 41.1636 | 0.00e+00  |
| '(loan_amnt) 14975<loan_amnt<=15000'                     | 0.5296854 | 0.0141852  | 37.3408 | 3.57e-305 |
| '(num_rev_tl_bal_gt_0) -Inf<num_rev_tl_bal_gt_0<=2'      | 0.4810422 | 0.0129686  | 37.0927 | 3.68e-301 |
| '(loan_amnt) 12000<loan_amnt<=14975'                     | 0.4583936 | 0.0130818  | 35.0405 | 5.44e-269 |



1 NAsThirdTraining <-
2   tibble(
3     name      = rownames(summary(SGLM_B_reretrain)$coefficient),
4     estimate  = summary(SGLM_B_reretrain)$coefficient$Estimate,
5     zValue    = abs(summary(SGLM_B_reretrain)$coefficient$"z value")
6   ) %>%
7   mutate(name = stringr::str_remove_all(name, "\`")) %>%
8   filter(name != "isGoodLoan") %>%
9
10  # Remove stray NAs (Shouldn't be any) or less than 2 sigmas
11  filter(is.na(estimate) | zValue < 2)
12
13 NAsThirdTraining <- NAsThirdTraining$name

speedglm:::summary.speedglm(SGLM_B_train) # AIC = 977037.2 # Log likelihood = -488350.6
speedglm:::summary.speedglm(SGLM_B_reretrain) # AIC = 982957 # Log likelihood = -491370.5
speedglm:::summary.speedglm(SGLM_B_reretrain) # AIC = 982957 # log Likelihood = -491370.5

```

3.5 Model result

3.5.1 Final list of variables

3.5.2 Scoring

Scoring expresses the coefficients that were estimated during the logistic regression into points on a scale. The conversion is done using three parameters that are chosen somewhat arbitrarily.

- the number of points increase / decrease that would reflect halving / doubling the odds of defaulting;
- an *anchoring* score reflecting a particular odd.

For our purpose, we will choose ($Score_{anchor}$) 5,000 points being equivalent to 1 in a 20 to default

$(Odds_{anchor})$, i.e. 5,000 points $\Leftrightarrow Odds_{Anchor} = \frac{1/20}{1-1/20}$. We will also choose 100 to reflect *times 2* change in odds (*DoubleOdds*). Those choices are completely arbitrary. Basically, the score is a linear representation of the odds. The score is defined by a point (the anchor) and the slope of the line going through that point.

Those values will reflect the total estimated score for a borrower (i.e. loan sample). The number of characteristics (information points gathered in a credit application), or number of bins, have to be irrelevant in calculating this score. In other words, if LendingClub were to gather 5 more information points, the score should, **mutatis mutandis**, be unchanged. (However, we would hope that the quality of the estimated score would improve.)

The score per variable needs to be adjusted using the number of information points, that is the number of characteristics. Here, the model has been trained on the number of bins. We first need to determine how many characteristics are used in the model.

```
1 ## [1] 24
```

We can now perform the scoring calculation.

Then we apportion across characteristics.

VERY IMPORTANT:

The intercept points have been allocated across all characteristics. Therefore the regression coefficient estimated for the intercept becomes redundant and needs removing.

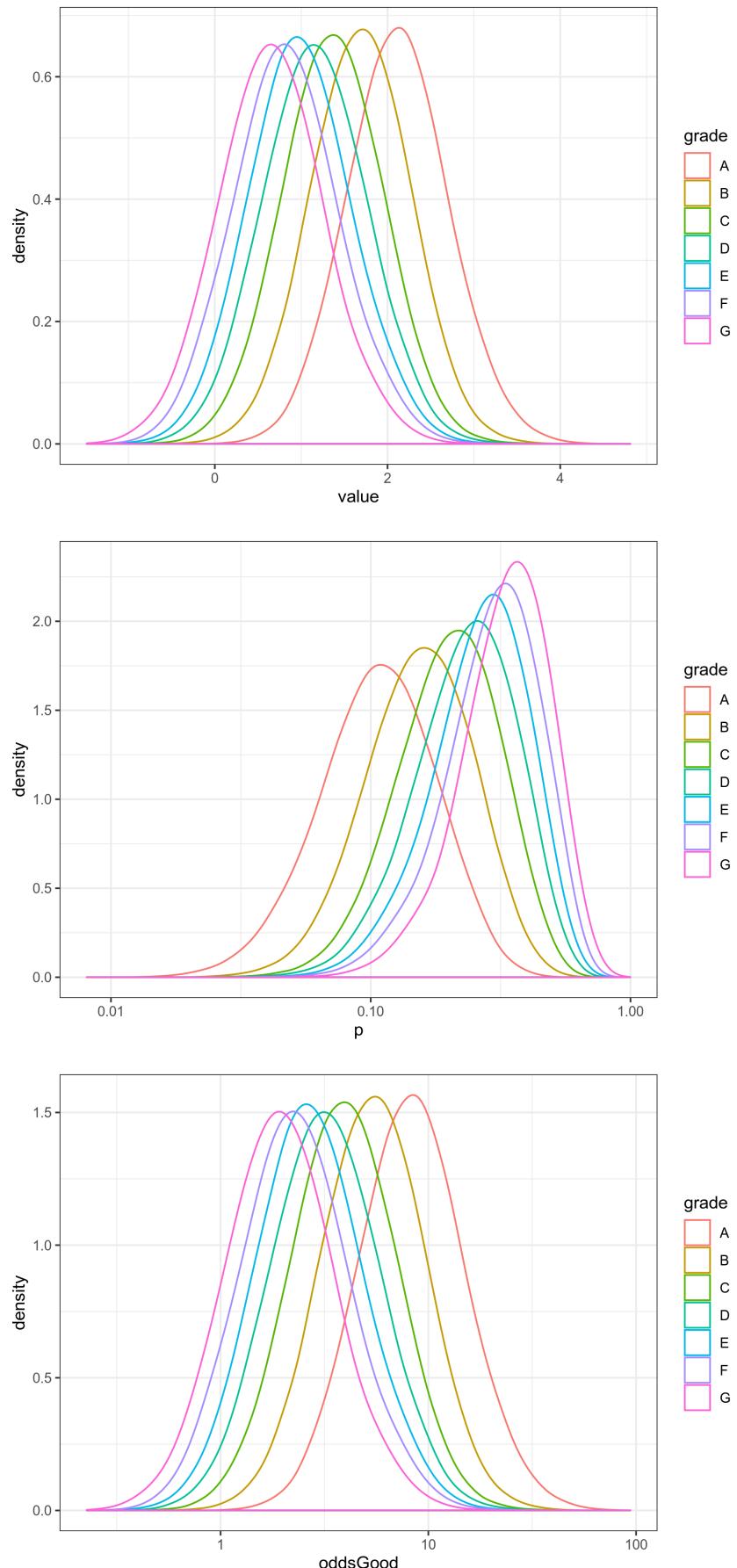
Using the model is a simple matrix multiplication: Loan matrix \times Scorecard weights

3.6 Training set

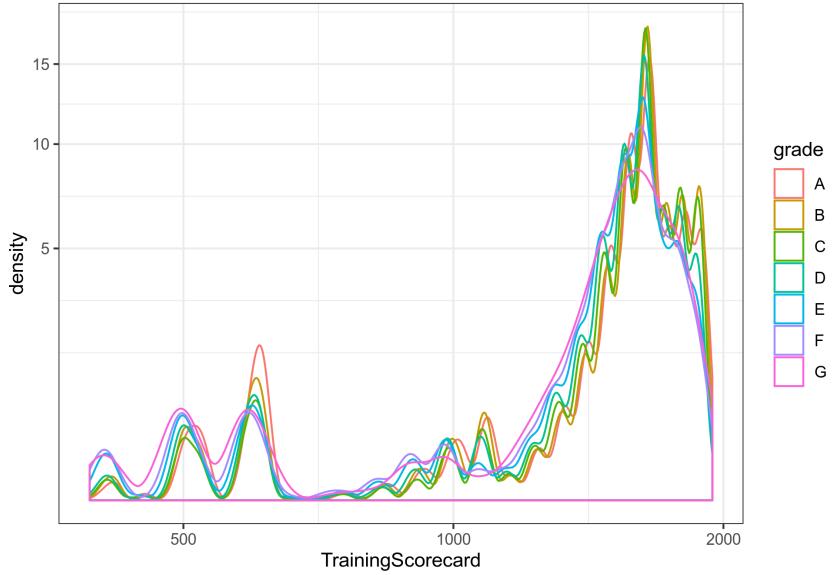
```
1 ##          used   (Mb) gc trigger   (Mb) max used   (Mb)
2 ## Ncells  2778407  148.4    4554342  243.3    4554342  243.3
3 ## Vcells 939305755 7166.4 1690871040 12900.4 1552998235 11848.5

1 ## [1] 1045084     168
1 ##          used   (Mb) gc trigger   (Mb) max used   (Mb)
2 ## Ncells  2778311  148.4    4554342  243.3    4554342  243.3
3 ## Vcells 707819569 5400.3 1690871040 12900.4 1552998235 11848.5
1 ## [1] 1045084     168
1 ## NULL
```

3.6.1 Histogram of the scores

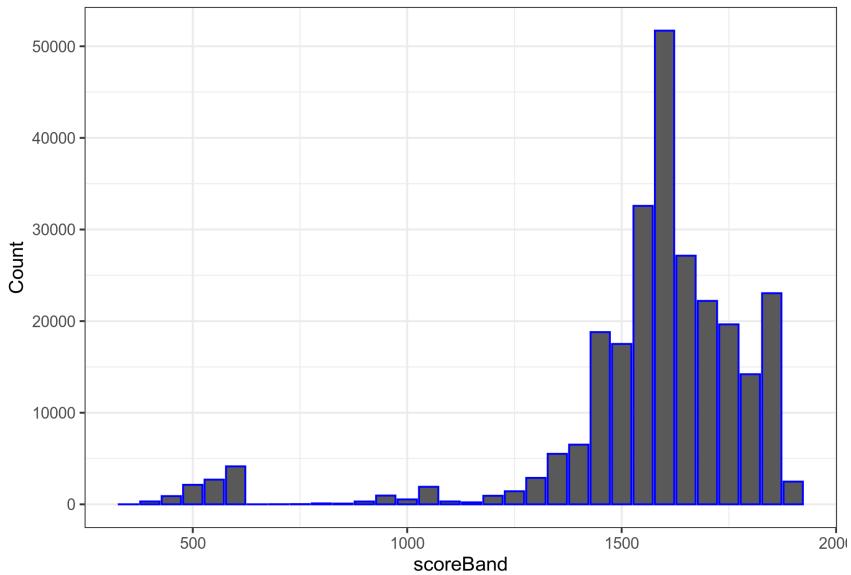


The following histogram shows the density plot of the scorecards. Note the log-scale for the score, and the square-root scale for the density. It mirrors the multimodal density distributions of previous visualisations.



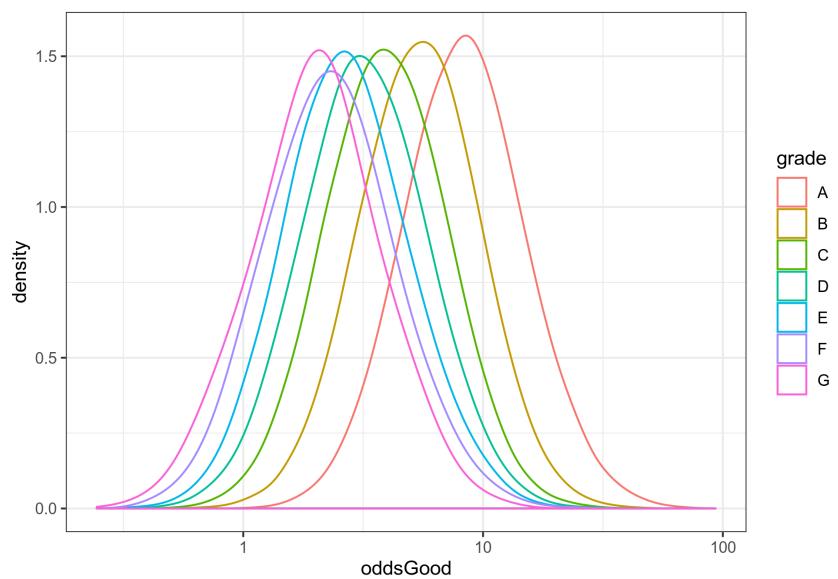
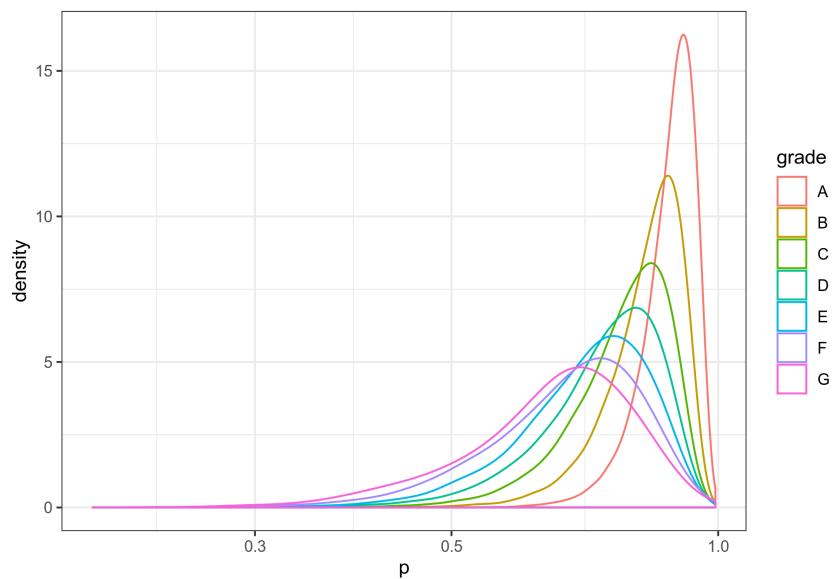
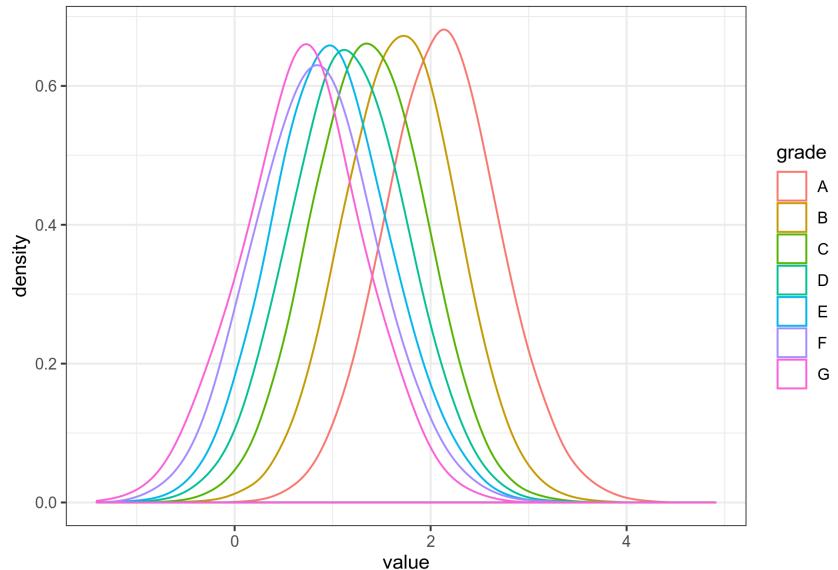
3.7 Test set

```
1 ## [1] 261272    222
1 ## [1] 261272    168
```



```
1 ## [1] 1636
```

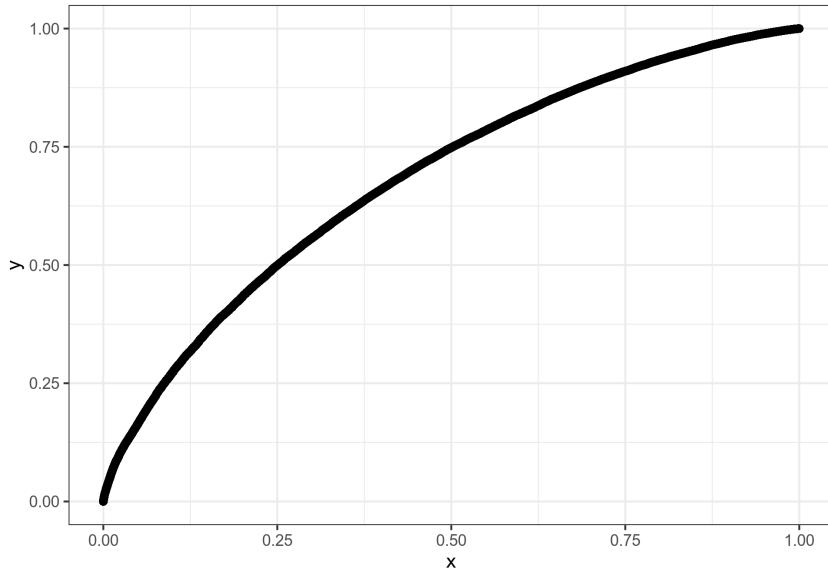
Same downward dynamics as training set



3.8 Correlation matrix

TODO

3.9 ROC Curve



```
1 ##  
2 ## Two-sample Kolmogorov-Smirnov test  
3 ##  
4 ## data: ROCRPlot$x and ROCRPlot$y  
5 ## D = 0.25527, p-value < 2.2e-16  
6 ## alternative hypothesis: two-sided  
  
1 ## Formal class 'performance' [package "ROCR"] with 6 slots  
2 ## ..@ x.name : chr "False positive rate"  
3 ## ..@ y.name : chr "True positive rate"  
4 ## ..@ alpha.name : chr "Cutoff"  
5 ## ..@ x.values :List of 1  
6 ## ...$ : num [1:255287] 0 0 0 0 0 0 0 0 0 0 ...  
7 ## ..@ y.values :List of 1  
8 ## ...$ : num [1:255287] 0.00 4.79e-06 9.58e-06 1.44e-05 1.92e-05 ...  
9 ## ..@ alpha.values:List of 1  
10 ## ...$ : num [1:255287] Inf 4.91 4.8 4.54 4.45 ...
```

3.10 Bin test set by prediction

```
1 ## Min. 1st Qu. Median Mean 3rd Qu. Max.  
2 ## 0.0000 0.5115 0.6030 0.5677 0.6655 0.8307
```

Chapter 4

Conclusion

Is the interest rate proposed by LC high enough?

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.

As a practical tool, this approach would not work in real life: we have no data on rejected loans. Using this model to accept/reject loans would need more work (using Reject Inference).

Currently only look at PD. Extend to LGD with Good Dollars and Bad Dollars instead of Good Loans / Bad Loans

Further possible explorations:

- Address the unbalanced dataset by sampling a number of training samples whose size is identical to the test dataset.
- Perform PCA beforehand
- Poisson GLM
- Improve the regularisation of the model parameters.
- Explore economic cycles/situations as additional entry.
- 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.
- Loss amount where good loan/bad loan proportion is replaced by a good dollar (repaid) / bad dollars (lost) fraction. See SAS Global Forum 2012.
- The size of the dataset is an issue to apply other techniques. But with stochastic methods, possible other models could be:
 - Tree models which have numerous variations (CART generally, and simple or aggregated boosted decision trees specifically)
 - Neural network

Chapter 5

Stochastic Gradient Descent

The diagram 5.1¹ shows a useful decision tree.

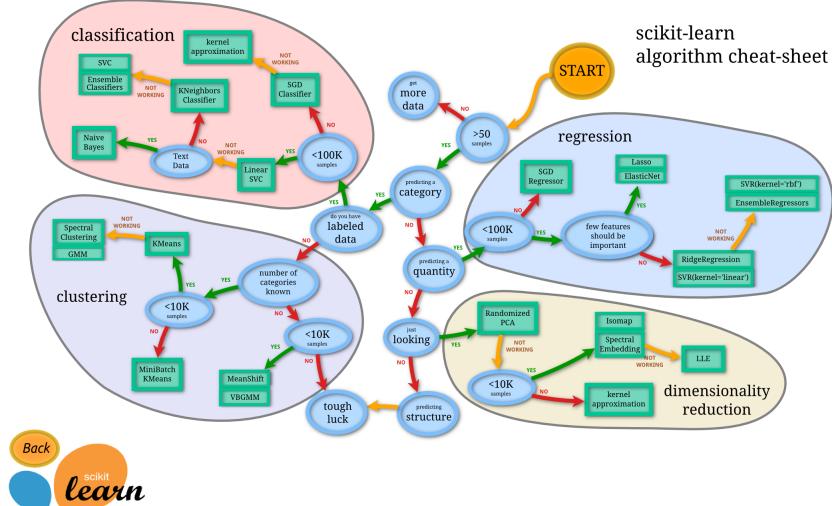


Figure 5.1: Scikit Learn algorithm cheat-sheet

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

5.1 Description of the model

5.2 Gradient descent

5.2.1 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:

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

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

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:

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

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

$$\nabla J(\theta) = (y^T - \theta X^T) X$$

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

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).

5.3 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.

5.3.0.1 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.

5.3.0.2 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 ².

5.3.1 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 $P$ an $Q$ 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`.
```

²<https://github.com/shabbychef/madness>

```

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$movieN,] <- spl_Q - alpha * delta_Q
77  }
78
79  # RMSE against the training set
80  error <- sqrt(sum(
81    tri_train$rating_z - rowSums(model$P[tri_train$userN,] *
82                                model$Q[tri_train$movieN,])
83  ) ^ 2)
84  / nSamples)
85
86  # Compares to RMSE before update
87  model$WORSE_RMSE <- (model$RMSE < error)
88  model$RMSE <- error
89
90  # Print some information to keep track of success
91  if (verbose) {
92    cat(
93      "  # features=",

```

```

94     ncol(model$P),
95     "  J=",
96     nSamples * error ^ 2 +
97       lambda / 2 * (sum(model$P ^ 2) + sum(model$Q ^ 2)),
98     "  Z-scores RMSE=",
99     model$RMSE,
100    "\n"
101  )
102  flush.console()
103 }
104
105 return(model)
106 }
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")

```

5.4 Various errands

5.4.1 Optimising the scorecard (or rating) depending on the NPV curve.

Input - Layer - Single Q - Distribution

Cost function as a function of the Q parameter (equivalent to scorecard).

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

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

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

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

We will center the distribution on the mode, therefore:

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

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 $\frac{\partial J}{\partial Q}$ is³:

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

³This was actually generated using Maxima (code in appendix) which allows for quicker iterations.

Appendix

5.5 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!

5.6 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*.

5.6.1 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  #
15  # STEP 2: Prepare the database as a tibble
16  #
17  #
18  ##
19  ## WARNING: THIS ASSUMES A 'datasets' DIRECTORY WAS CREATED
20  ##
21  library(RSQLite)
22  db_conn <-
23    dbConnect(RSQLite::SQLite(), "datasets/lending_club.sqlite")
24  dbListTables(db_conn)
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 }

```

5.6.2 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*⁴ 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 })

```

5.6.3 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>

⁴<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")) %/%% days(1)
44   RATES <- tibble(n = seq(0, 7212)) %>%
45
46     # Create a column with all dates
47     mutate(DATE = ymd("2000-01-01") + days(n)) %>%
48     select(-n) %>%
49
50     # Add all daily 3- then 5-year rates and fill missing down
51     left_join(RATES3Y) %>%
52     fill(RATE3Y, .direction = "down") %>%
53

```

```

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

```

5.6.4 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   ##
4   ## Median income per household by FIPS from 2002 to 2017
5   ##
6   # Prepare median income
7   medianIncome <-
8   # Load the dataset after dropping the first line
9   read.csv(
10    "datasets/csv/GeoFRED_Estimate_of_Median_Household_Income_by_County_Dollars.csv",
11    sep = "\t",
12    skip = 1,
13    stringsAsFactors = FALSE
14  ) %>%
15
16  # Drops columnsn containing a unique identifier and the FIPS name
17  select(-"Series.ID", -"Region.Name") %>%
18
19  # Rename the relevant column to 'FIPS'
20  rename(FIPS = "Region.Code") %>%
21
22  # Order by FIPS
23  arrange(FIPS) %>%
24
25  # Convert to a 'long' table, i.e. one column for FIPS, one for date, one for income
26  pivot_longer(cols = starts_with("X"),
27    names_to = "Date",
28    values_to = "medianIncome") %>%
29
30  # Create actual dates
31  mutate(Date = str_replace(Date, "[X]", ""))

```

```

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 }

```

5.7 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

5.8 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).

5.8.1 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 = 1000,
21                           intRate = 0.02,
22                           term = 36,
23                           totalPaid = 1000,
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 late 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 }
148
149 #  

150 # Calculate all the IRRs and month of default for all the loans.  

151 # WARNING: This takes around a full day to run!!!!  

152 #
153 # The actual data was generated by the Julia version, with cross-checks.  

154 # Julia version takes about 150 sec on the same unoptimised code.  

155 #
156 local({
157     loansIRR <-
158     loans %>%
159     rowwise() %>%
160     do(
161         calculateIRR(
162             loanNumber = .$loanID,
163             loan = .$funded_amnt,
164             intRate = .$int_rate,
165             term = .$term,
166             totalPaid = .$total_pymnt,
167             totalPrincipalPaid = .$total_rec_prncp,
168             totalInterestPaid = .$total_rec_int,
169             recoveries = .$recoveries,
170             lateFees = .$total_rec_late_fee
171         )
172     )
173
174     saveRDS(loansIRR, "datasets/lending_club_IRRs.rds")
175
176 })

```

5.8.2 Julia code

5.8.2.1 Internal Rate of Return

5.8.3 Maxima derivation of the cost function

$\text{PDF1}(x, Q) := \alpha_1(Q) * \sqrt{1 / (2 * \pi)} * \exp(-x^2 / (2 * Q))$

```

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) ;

```

5.9 System version

1 ##	sysname	release
2 ##	"Linux"	"5.3.0-24-generic"
3 ##	version	nodename
4 ## "#26-Ubuntu SMP Thu Nov 14 01:33:18 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.