P2P Loan Performance on Lending Club

Peter Jin
phj@cs.berkeley.edu

November 25, 2014

Objectives

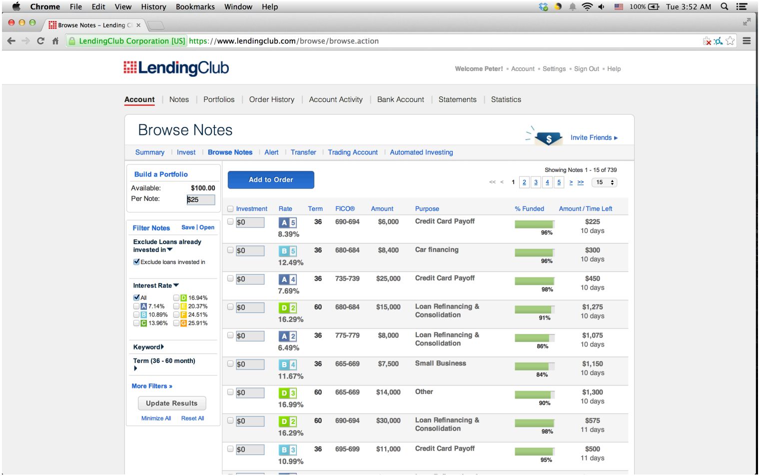
My questions to you:

1. Did I skip over some background knowledge?
2. What other plots am I missing and should add?
3. How’s my ~~driving~~ methodology?

Background

- Individual borrowers with Internet access apply for an uncollateralized loan on a P2P lending platform (Lending Club, Prosper).
- Individual investors can fund parts of other individuals’ loans through the same platform.
- The platform takes a cut of the loan payments.

Background



Chrome File Edit View History Bookmarks Window Help
100% Tue 3:52 AM
Browse Notes – Lending C
LendingClub Corporation [US] https://www.lendingclub.com/browse/browse.action
LendingClub
Welcome Peter! • Account • Settings • Sign Out • Help
Account | Notes | Portfolios | Order History | Account Activity | Bank Account | Statements | Statistics
Browse Notes
Invite Friends
Summary | Invest | Browse Notes | Alert | Transfer | Trading Account | Automated Investing
Build a Portfolio
Available: $100.00
Per Note: $25
Add to Order
Showing Notes 1 - 15 of 739
Filter Notes Save | Open
Exclude Loans already invested in
Exclude loans invested in
Interest Rate
All
A 7.14%
B 10.89%
C 13.96%
D 16.94%
E 20.37%
F 24.51%
G 25.91%
Keyword
Term (36 - 60 month)
More Filters »
Update Results
Minimize All Reset All
Investment Rate Term FICO® Amount Purpose % Funded Amount / Time Left
$0 A5 8.39% 36 690-694 $6,000 Credit Card Payoff 96% $225 10 days
$0 B5 12.49% 36 680-684 $8,400 Car financing 96% $300 10 days
$0 A4 7.69% 36 735-739 $25,000 Credit Card Payoff 98% $450 10 days
$0 D2 16.29% 60 680-684 $15,000 Loan Refinancing & Consolidation 91% $1,275 10 days
$0 A2 6.49% 36 775-779 $8,000 Loan Refinancing & Consolidation 86% $1,075 10 days
$0 B4 11.67% 36 665-669 $7,500 Small Business 84% $1,150 10 days
$0 D3 16.99% 60 665-669 $14,000 Other 90% $1,300 10 days
$0 D2 16.29% 60 690-694 $30,000 Loan Refinancing & Consolidation 98% $575 11 days
$0 B3 10.99% 36 695-699 $11,000 Credit Card Payoff 95% $500 11 days

Background

The goal of an investor is to turn a profit. To do so requires a correct valuation of a loan. One (simplified) method of valuation is the expected discounted cashflows:

$$V(x) = \sum_{k=1}^{K} \frac{iP(T \geq k|x)}{(1+\gamma)^k}$$

where K is the term of the loan in months, i is the net monthly installment (after fees), $P(T \geq k|x)$ is the probability that the loan with feature vector x makes at least k payments, and $\gamma \geq 0$ is a discount rate (takes into account the time-value of money).

Analysis Targets

1. Define and characterize loan durations before default and prepayment.
2. How do loan durations differ based on their features?
3. How does the addition of a dataset change or augment our analysis?

The Data

Two datasets:

1. Dataset 1: Snapshots of historical loan issues from June 2007 to June 2014, with loan info, loan status, and borrower credit profile. This is updated quarterly, and is the main public dataset distributed by Lending Club.
2. Dataset 2: Detailed payment histories for each loan, as well as the evolving credit profile of the borrower. This was recently released by Lending Club (up-to-date as of 11/7) and is tucked away in a corner of their website.

Dataset 1

- CSV format with 100 fields. Newest version (2014Q3) has only 56 fields (non-members see 52 fields, where the missing 4 fields are credit scores).
- A handful of data-munging issues (extraneous line breaks and comments), but generally without problems.
- Has information like: loan ID, borrower ID, loan amount, term, grade, interest rate, borrower city, income, credit score, detailed credit profile, last payment date, cumulative payments…

Dataset 1

```
"id","member_id","loan_amnt","funded_amnt","funded_amnt_inv","term","int_rate",
"installment","grade","sub_grade","emp_title","emp_length","home_ownership",
"annual_inc","is_inc_v","accept_d","exp_d","list_d","issue_d","loan_status",
"pymnt_plan","url","desc","purpose","title","addr_city","addr_state",
"acc_now_delinq","acc_open_past_24mths","bc_open_to_buy","percent_bc_gt_75",
"bc_util","dti","delinq_2yrs","delinq_amnt","earliest_cr_line",
"fico_range_low","fico_range_high","inq_last_6mths",
"mths_since_last_delinq","mths_since_last_record","mths_since_recent_inq",
"mths_since_recent_revol_delinq","mths_since_recent_bc","mort_acc","open_acc",
"pub_rec","total_bal_ex_mort","revol_bal","revol_util","total_bc_limit",
"total_acc","initial_list_status","out_prncp","out_prncp_inv","total_pymnt",
"total_pymnt_inv","total_rec_prncp","total_rec_int","total_rec_late_fee",
"recoveries","collection_recovery_fee","last_pymnt_d","last_pymnt_amnt",
"next_pymnt_d","last_credit_pull_d","last_fico_range_high","last_fico_range_low",
"total_il_high_credit_limit","num_rev_accts","mths_since_recent_bc_dlq",
"pub_rec_bankruptcies","num_accts_ever_120_pd","chargeoff_within_12_mths",
"collections_12_mths_ex_med","tax_liens","mths_since_last_major_derog",
"num_sats","num_tl_op_past_12m","mo_sin_rcnt_tl","tot_hi_cred_lim","tot_cur_bal",
"avg_cur_bal","num_bc_tl","num_actv_bc_tl","num_bc_sats","pct_tl_nvr_dlq",
"num_tl_90g_dpd_24m","num_tl_30dpd","num_tl_120dpd_2m","num_il_tl",
"mo_sin_old_il_acct","num_actv_rev_tl","mo_sin_old_rev_tl_op",
"mo_sin_rcnt_rev_tl_op","total_rev_hi_lim","num_rev_tl_bal_gt_0","num_op_rev_tl",
"tot_coll_amt","policy_code"
```

Dataset 1

”54734”,”80364”,”25000”,”25000”,”19080.057198275422”,” 36 months”,” 11.89%”,
”829.1”,”B”,”B4”,””,”< 1 year”,”RENT”,”85000”,”Verified”,
”2009-07-26”,”2009-08-09”,”2009-07-26”,”2009-08-05”,”Fully Paid”,”n”,
”https://www.lendingclub.com/browse/loanDetail.action?loan_id=54734”,
”Due to a lack of personal finance education and exposure to poor financing skills
growing up, I was easy prey for credit predators. I am devoted to becoming debt-free
and can assure my lenders that I will pay on-time every time. I have never missed a
payment during the last 16 years that I have had credit. ”,”debt_consolidation”,”Debt
consolidation for on-time payer”,
”San Francisco”,”CA”,”0”,””,””,””,””,”19.48”,”0”,”0”,”1994-02-15 10:39”,
”735”,”739”,”0”,””,””,””,””,””,””,”10”,”0”,””,”28854”,”52.1%”,””,”42”,”f”,
”0.00”,”0.00”,”29324.32”,”21811.70”,”25000.00”,”4324.32”,”0.0”,”0.0”,”0.0”,
”2011-10-14”,”7392.08”,”null”,”2012-08-28”,”789”,”785”,””,””,””,”0”,””,”0”,
”0”,”0”,”1”

Dataset 1


```
”10158748”,”12010420”,”12000”,”12000”,”12000”,” 60 months”,” 14.47%”,
”282.16”,”C”,”C2”,”Clerk”,”10+ years”,”RENT”,”48000”,”Verified”,
”2013-12-29”,”2014-01-12”,”2013-12-30”,”2013-12-31”,”Charged Off”,”n”,
”https://www.lendingclub.com/browse/loanDetail.action?loan_id=10158748”,
” Borrower added on 12/29/13 > Pay off Credit Cards<br><br> Borrower added on 12/29/13 >
payoff credit cards<br>”,”credit_card”,”Consolidate”,
”REDDING”,”CA”,”0”,”6”,”1581”,”50”,”65.6”,”18.6”,”0”,”0”,”2000-06-29 12:00”,
”675”,”679”,”0”,””,”113”,”8”,””,”20”,”0”,”15”,”1”,”56182”,”3576”,”65%”,”4600”,
”24”,”f”,”0.00”,”0.00”,”1127.27”,”1127.27”,”559.20”,”568.07”,”0.0”,”0.0”,”0.0”,
”2014-05-06”,”282.16”,””,”2014-07-22”,”599”,”595”,”47504”,”8”,””,”1”,”0”,”0”,
”0”,”0”,””,”15”,”3”,”8”,”53004”,”56182”,”4013”,”6”,”1”,”2”,”100”,”0”,”0”,”0”,
”16”,”79”,”2”,”162”,”8”,”5500”,”2”,”4”,”0”,”1”
```

Dataset 1

”20282255”,””,”10976.0”,”10976.0”,”10976.0”,””,””,””,””,””,””,””,””,””,””,””,
””,””,”2014-05-27”,””,””,””,””,””,””,”San Francisco”,”CA”,””,””,””,””,””,””,
””,
””,
””,””,””,””,””,””,””,””,””,””,””,””,””,””,”2”

Dataset 1

What's this "`policy_code`" = "`2`" all about?
From a third party blog post on 11/21/13:

- "These are loans made to borrowers that do not meet Lending Club's current credit policy standards."
- "The FICO scores on these borrowers are typically 640-659, below the 660 threshold on Policy Code 1 loans."
- **"These loans are made available to select institutional investors who have a great deal of experience with consumer loans in this credit spectrum and with Lending Club."**
- "Lending Club believes that Policy 2 loans could grow to **a total of 15% of the total volume over the next 12 months**."

Dataset 2

- CSV format with 39 fields.
- Two files: one with net payments, the other with the payments allocated for investors.
- Loan payment history cross references the loan ID from dataset 1.

Dataset 2

```
LOAN_ID,RECEIVED_D,PERIOD_END_LSTAT,Month,MOB,CO,PBAL_BEG_PERIOD_INVESTORS,
PRNCP_PAID_INVESTORS,INT_PAID_INVESTORS,FEE_PAID_INVESTORS,DUE_AMT_INVESTORS,
RECEIVED_AMT_INVESTORS,PBAL_END_PERIOD_INVESTORS,MONTHLYPAYMENT_INVESTORS,
COAMT_INVESTORS,InterestRate,IssuedDate,dti,State,HomeOwnership,MonthlyIncome,
EarliestCREDITLine,OpenCREDITLines,TotalCREDITLines,RevolvingCREDITBalance,
RevolvingLineUtilization,Inquiries6M,DQ2yrs,MonthsSinceDQ,PublicRec,
MonthsSinceLastRec,EmploymentLength,currentpolicy,grade,term,appl_fico_band,
vintage,PCO_RECOVERY_INVESTORS,PCO_COLLECTION_FEE_INVESTORS
```

Dataset 2

54734,SEP09,Current,SEP09,1,0,19080.0572,443.64790001,189.12311697,0,
632.77101698,632.77101698,18636.4093,632.77101698,0,0.118900,AUG09,19.48,CA,
RENT,7083.3333333,FEB94,10,42,28854,0.521,0,0„0„< 1 year,1,B,36,735–739,09Q3„

54734,OCT09,Current,OCT09,2,0,18636.4093,448.04537497,184.72564202,0,
632.77101698,632.77101698,18188.363925,632.77101698,0,0.118900,AUG09,19.48,CA,
RENT,7083.3333333,FEB94,10,42,28854,0.521,0,0„0„< 1 year,1,B,36,735–739,09Q3„

⋮

54734,SEP11,Current,SEP11,25,0,6187.9023026,623.70010638,61.335003282,0,
632.77101698,685.03510966,5564.2021962,632.77101698,0,0.118900,AUG09,19.48,CA,
RENT,7083.3333333,FEB94,10,42,28854,0.521,0,0„0„< 1 year,1,B,36,735–739,09Q3„

54734,OCT11,Fully Paid,OCT11,26,0,5564.2021962,5586.4995332,55.152835853,0,
632.77101698,5641.6523691,0,632.77101698,0,0.118900,AUG09,19.48,CA,RENT,
7083.3333333,FEB94,10,42,28854,0.521,0,0„0„< 1 year,1,B,36,735–739,09Q3„

Dataset 2

My main gripe with dataset 2 is that it no longer lists exact dates of origination/payment, but instead it bins loans by their monthly cohort. (Worse still, Lending Club did the same thing to the newest version of dataset 1, released concurrently.)

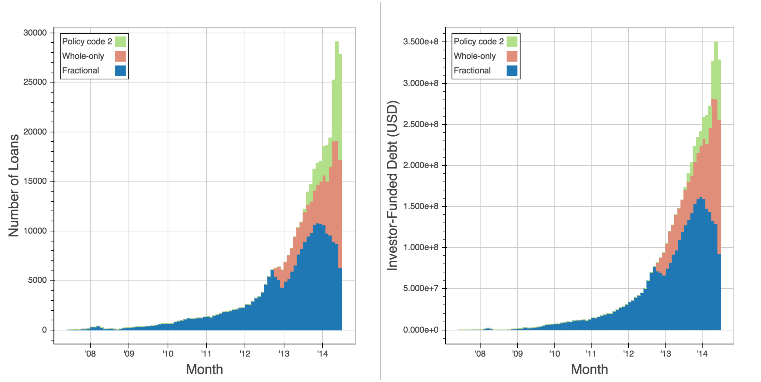
This decreases the resolution of the data and is generally annoying, but I can live with it.

The Data

What parts of the data do we care about?

- In general, we use the “investors” version of numbers, since that is what users of Lending Club will see.
- We already saw that “policy code” 2 loans are a nonstarter. This leaves 87.9% of the original data.
- The remaining loans can be “fractional” or “initially whole-loan-only.” The latter are typically invested in by institutional investors. This is 73.4% of the remaining data, or 64.6% of the original data.
- The most recent data (first 6 months of 2014) throw off some of the statistical estimates because their payment histories are too short. Omitting them leaves 50.3% of the original data.

The Data



Policy code 2
Whole-only
Fractional
Number of Loans
30000
25000
20000
15000
10000
5000
0
'08
'09
'10
'11
'12
'13
'14
Month
Policy code 2
Whole-only
Fractional
Investor-Funded Debt (USD)
3.500e+8
3.000e+8
2.500e+8
2.000e+8
1.500e+8
1.000e+8
5.000e+7
0.000e+0
'08
'09
'10
'11
'12
'13
'14
Month

Loan Durations

What’s a loan duration, anyway? We care about four events:

1. A loan is paid off on time.
2. A loan is fully paid off but late.
3. A loan is fully paid off early.
4. A loan is never fully paid off (charged off).

Chargeoff is the most pernicious event. Without further qualification, “loan duration” will refer to loan duration before chargeoff.

Loan Durations

The data for a loan tells us:

- The date the loan was issued/originated.
- The total funded amount (due to investors) on the loan.
- The installment on the loan.
- The date of the borrower's last payment on the loan.
- The total amount paid on the loan by the borrower.

Loan Durations

We can come up with at least two definitions for "loan duration":

1. The number of days between the issue date and the date of last payment.
2. An approximation for the number of payments; namely, the minimum of:
   (a) the previous definition in units of months;
   (b) the ratio of the amount paid by the loan installment.

Loan Durations

Assumptions and acceptable conditions for the second definition:

1. A borrower makes an unbroken sequence of monthly payments, then stops either due to default, prepayment, or maturity of the loan.
2. The definition is conservative, in the sense that any deviation in the actual payment history results in slightly more interest paid (on the remaining principal).

Censorship

In an observational study, subjects/data are *censored* when the study ends before the random event of interest (e.g., chargeoff or prepayment) can be observed.

Censorship

Define the *survival function* $S(t)$ as the probability that the subject's observed duration T until an event, which is a random variable, is greater than t. In other words, $S(t) = P(T > t)$.

If the cumulative distribution function of subject durations is $F(t)$, then $S(t) = 1 - F(t)$.

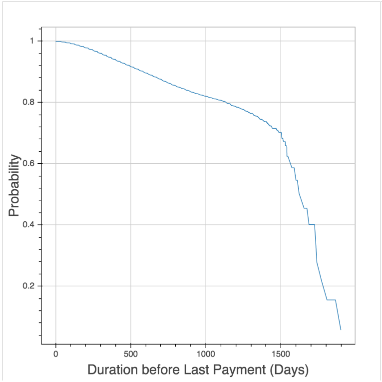
Censorship

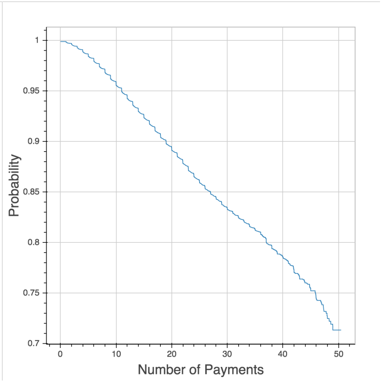
Two main estimators for $S(t)$:

1. Kaplan-Meier estimate, $\hat{S}(t)$.
2. Nelson-Aalen estimate of the *cumulative hazard function*, $\hat{\Lambda}(t)$, and transforming $\tilde{S}(t) = e^{-\hat{\Lambda}(t)}$.

Nelson-Aalen is typically greater than or equal to Kaplan-Meier, which can go to zero at the rightmost end.

Distribution of Loan Durations





Distribution of Loan Durations

There are two terms of loans:

- 36 months (1095 days);
- 60 months (1825 days).

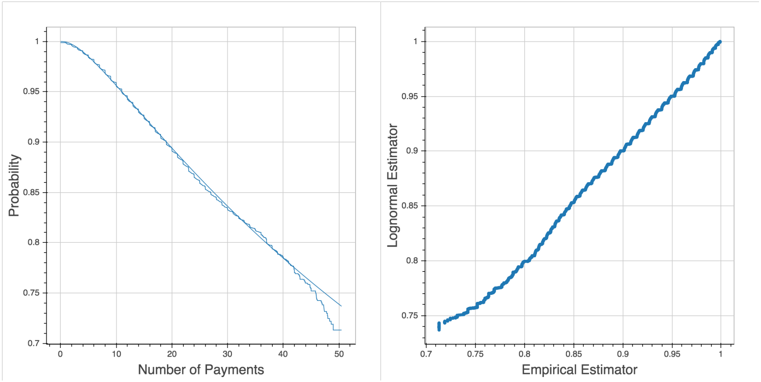
Distribution of Loan Durations

Common parametric survival functions:

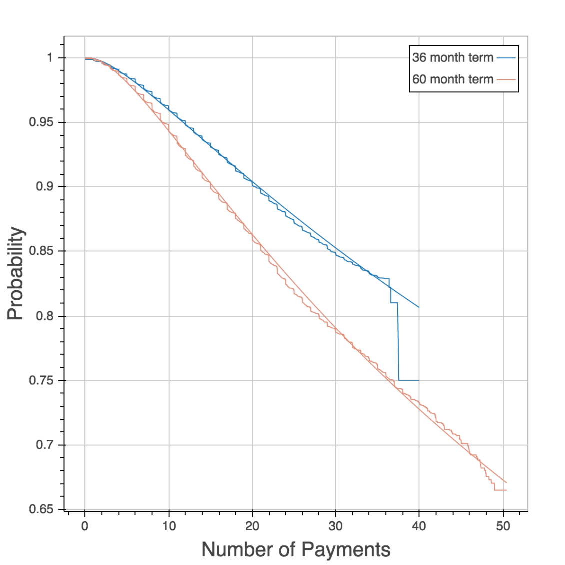
- Exponential;
- Weibull;
- Log-logistic;
- Lognormal.

We pick a lognormal distribution to fit loan durations until chargeoff.

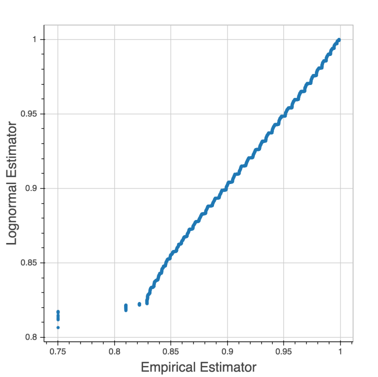
Distribution of Loan Durations

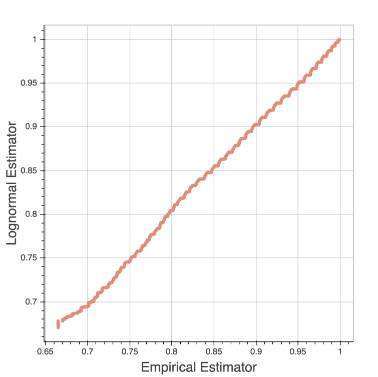


Distribution of Loan Durations



Distribution of Loan Durations

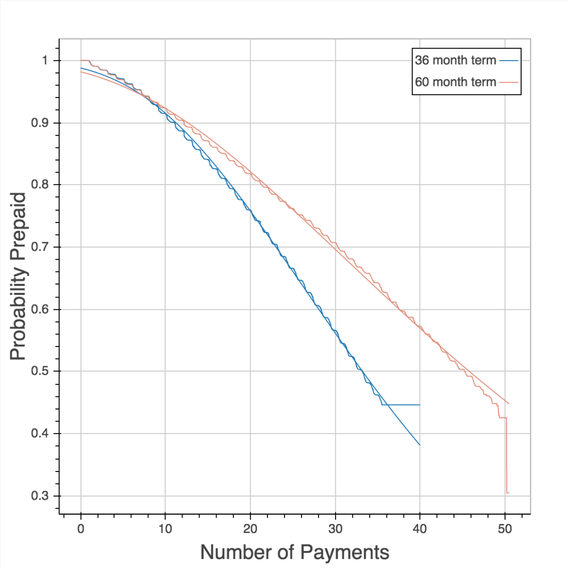




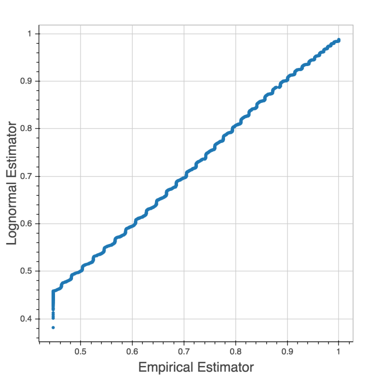
Distribution of Loan Durations

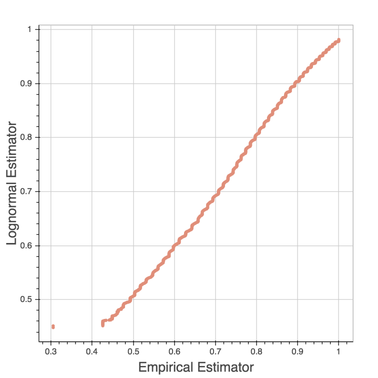
We also use a lognormal distribution to fit loan durations until prepayment, but we needed to use a location parameter.

Distribution of Loan Durations



Distribution of Loan Durations





Features

A graphical way for comparing the effect of a covariate or a feature on the survival function:

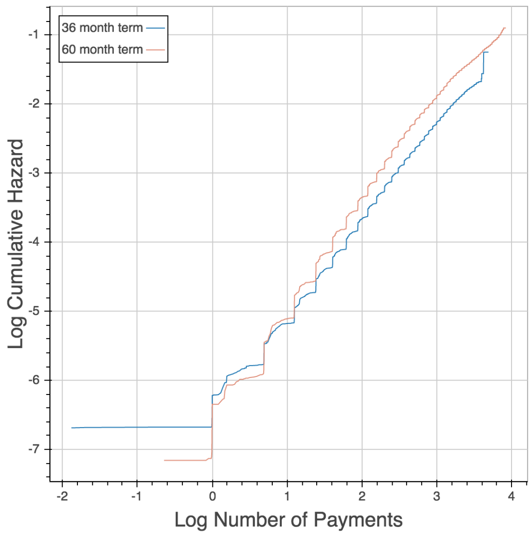
- Plot the cumulative hazard $\Lambda(t) = -\log(S(t))$ over t in log-log scale.
- The two curves are horizontally shifted $\Rightarrow$ “accelerated failure time.”
- The two curves are vertically shifted $\Rightarrow$ “proportional hazards.”

Features

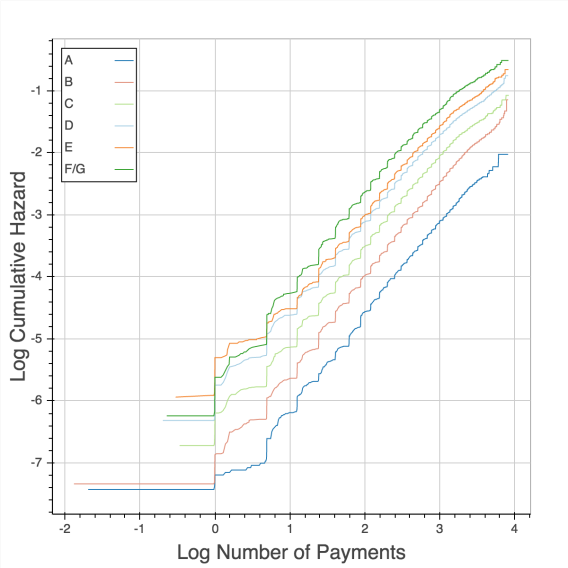
Things not talked about:

- Log rank test;
- Parametric accelerated failure time regression;
- Semiparametric (Cox) proportional hazards regression.

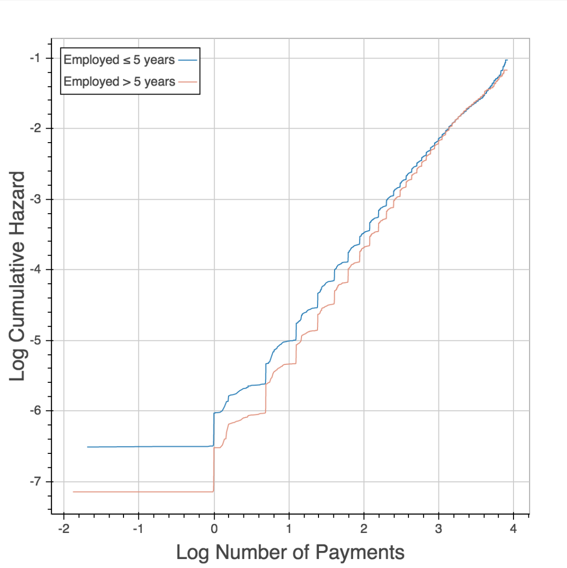
Features


36 month term
60 month term
Log Cumulative Hazard
Log Number of Payments
-1
-2
-3
-4
-5
-6
-7
-2
-1
0
1
2
3
4

Features

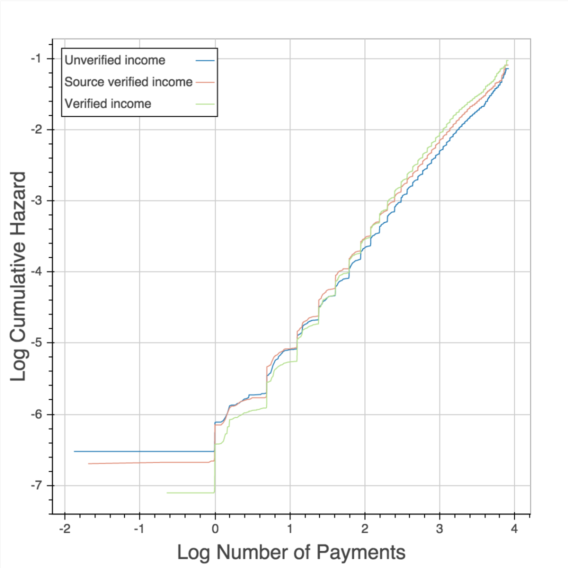

A
B
C
D
E
F/G
Log Cumulative Hazard
-1
-2
-3
-4
-5
-6
-7
-2
-1
0
1
2
3
4
Log Number of Payments

Features



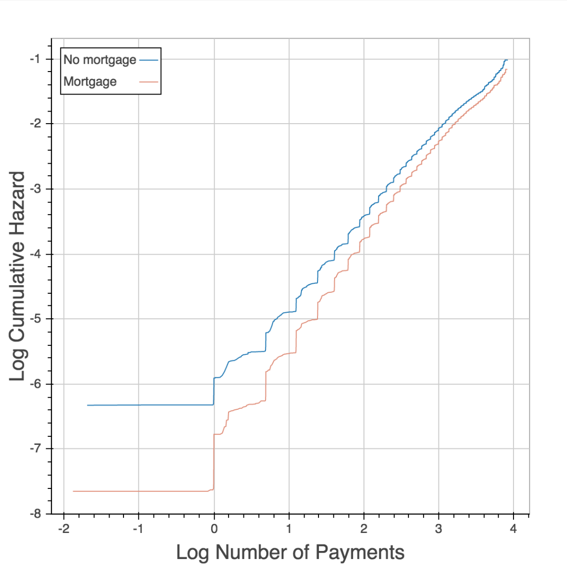
Employed ≤ 5 years
Employed > 5 years
-1
-2
-3
-4
-5
-6
-7
Log Cumulative Hazard
-2
-1
0
1
2
3
4
Log Number of Payments

Features



Unverified income
Source verified income
Verified income
Log Cumulative Hazard
Log Number of Payments
-1
-2
-3
-4
-5
-6
-7
-2
-1
0
1
2
3
4

Features

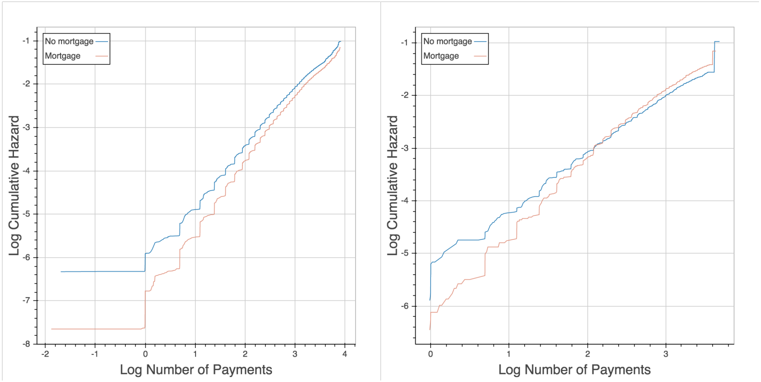

No mortgage
Mortgage
-1
-2
-3
-4
-5
-6
-7
-8
Log Cumulative Hazard
-2
-1
0
1
2
3
4
Log Number of Payments

Nonstationarity

Effects of features change in time. A good example of this:

- Subprime financial crisis starts in summer 2007.
- Coincidentally, Lending Club issued their first loan in June 2007.
- Consider loans with and without mortgages issued from 2007-2009, and compare to the whole dataset.

Nonstationarity



Late Payments

In fact, our mental model of loan durations is oversimplified.

Loans can transition through various stages of lateness before charging off:

- a grace period (no penalty for being 1-15 days late);
- 16-30 days late;
- 31-120 days late;
- default (121-150 days late).

These are specific categories given to us by Lending Club.

Late Payments

Further caveats:

- Dataset 1 only shows the lateness status for loans which are currently late (at the time the dataset was prepared).
- Dataset 1 also shows the late fees paid, but not when or for how long a borrower was late paying.
- As an approximation, we upgrade all late loans into charged off loans. This is strictly not correct. Lending Club reports the following percentages of late loans eventually becoming “net charged off”:
  1. 23% of loans in the grace period (1-15 days late);
  2. 58% of loans 16-30 days late;
  3. 75% of loans 31-120 days late;
  4. and 91% of defaulted loans.

  “Net charged off” loans are a subset of all charged off loans.

Late Payments

Luckily, we have Dataset 2 to get to the bottom of this question.

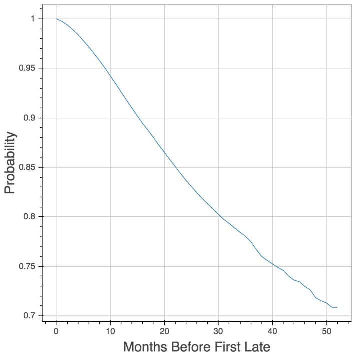
- After pruning loans that are policy code 2, whole-only, or issued in 2014, we have a total of 190851 loans from Dataset 1.
- Of the remainder, 190618 (99.9%) can be cross-referenced with Dataset 2.

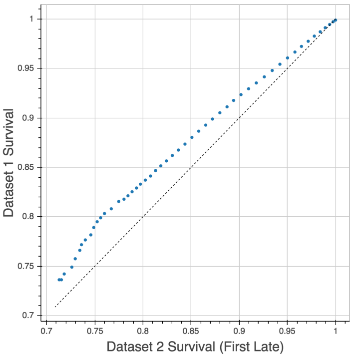
Late Payments

We can track at least two interesting events that happen:

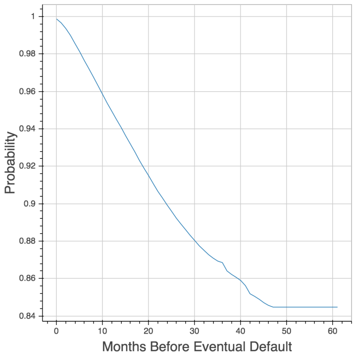
- Months paid before first late (non-)payment.
- Months paid before eventual default/chargeoff.

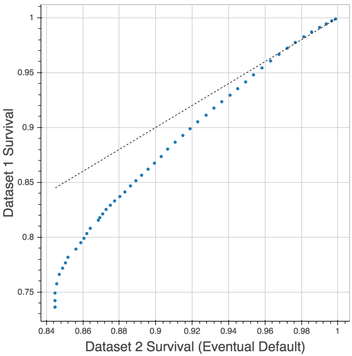
Late Payments



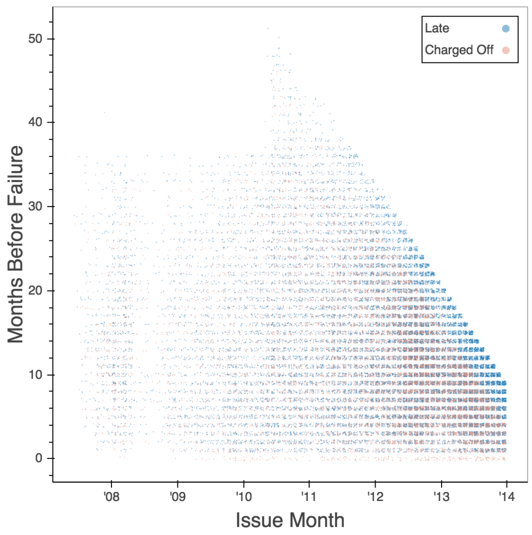


Late Payments

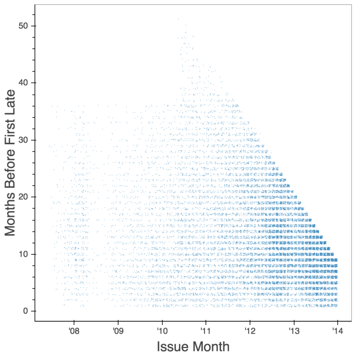


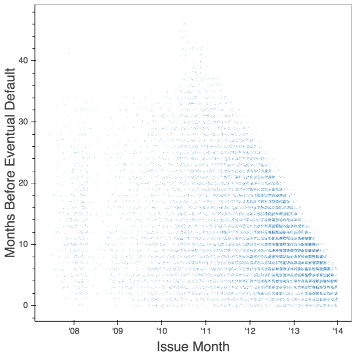


Late Payments

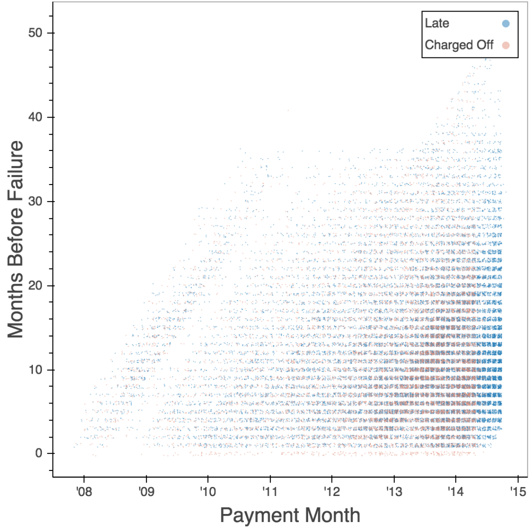


Late Payments

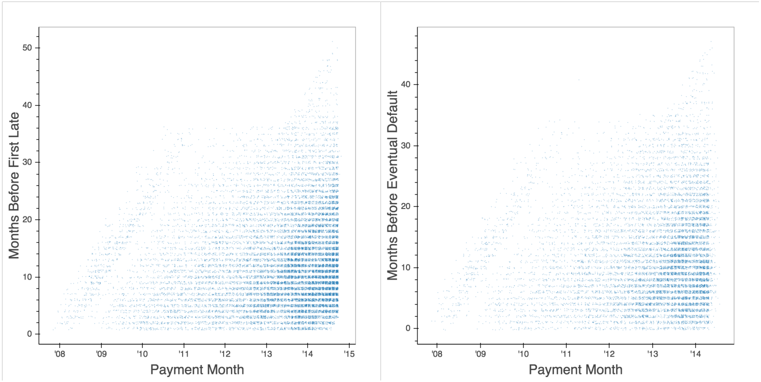




Late Payments



Late Payments



Late Payments

Still more work to be done!

Summary

- Characterized the distribution of a censored dataset.
- Inspected the effects of covariates/features.
- Revisited previous results using new data.

Questions?