ECO6936: Data Appendix

Proposed Title: Modeling default probabilities in peer-to-peer services

Nigar Sultana

07/10/2020

1 Data

Lending Club, founded in 2007, is the world's largest P2P lending platform with over 20 billion USD in loan issuance. It offers consumer and small- and medium-sized enterprise (SME) loans over fixed periods of 36 or 60 months. They provide detailed public information about each available loan, and allow downloading historical information of all the loans funded (see https://www.lendingclub.com/info/download-data.action). I have used the data frame from 2012-2015 for my analysis.

1.1 Data Description

The borrowers' loan data from 2012-2015 were recorded in CSV format per year. The data frame consists of loan history of loan applications recorded in the *Lending Club* system between the time period January 2012 and December 2015. The data comprises of demographic

and financial information of borrowers and the corresponding loan transactions. A data dictionary is provided in a separate file in the data frame. In total, there were three data frames. Analysis was conducted in R (Team et al., 2013). First, I have imported the data frames in R and combined them together. Initially, the data frame had 844,907 rows and 150 variables overall. The data frame has included a larger number of missing observations, wrong format of the variables, and several blank columns. The comprehensive data wrangling procedure is described in the Data Appendix section.

After cleaning the data frame, I have found three types of variables: numerical, categorical, and date variables. The description of the variables are given in table 1, 2, and 3.

Numerical Variables:

Table 1: Description of Numerical Variables

Numerical Variables	Description
acc-open-past-24mths	Number of trades opened in past 24 months.
annual-inc	The self-reported annual income provided by the borrower during registration.
acc-now-delinq	The number of accounts on which the borrower is now delinquent.
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-2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
delinq-amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.

Table 1: Description of Numerical Variables

Numerical Variables	Description
dti	A ratio calculated using the borrowers total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrowers self-reported monthly income.
delinq-2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
funded-amnt-inv	The total amount committed by investors for that loan at that point in time.
id	A unique LC assigned ID for the loan listing.
int-rate	Interest Rate on the loan
last-fico-range-high	The upper boundary range the borrowers last FICO pulled belongs to.
last-fico-range-low	The lower boundary range the borrowers last FICO pulled belongs to.
last-pymnt-amnt	Last total payment amount received
loan-amnt	The listed amount of the loan applied for by the borrower.
mo-sin-old-il-acct	Months since oldest bank installment 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-last-delinq	The number of months since the borrower's last delinquency.
mths-since-last-major-derog	Months since most recent 90-day or worse rating
mths-since-recent-bc	Months since most recent bankcard account opened.
mths-since-recent-inq	Months since most recent inquiry.

Table 1: Description of Numerical Variables

Numerical Variables	Description
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
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-tl-120dpd-2m	Number of accounts currently 120 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-30dpd	Number of accounts currently 30 days past due (updated in past 2 months)
num-tl-op-past-12m	Number of accounts opened in past 12 months
open-acc	The number of open credit lines in the borrower's credit file.
pct-tl-nvr-dlq	Percent of trades never delinquent
percent-bc-gt-75	Percentage of all bankcard accounts >75% of limit.
pub-rec	Number of derogatory public records
pub-rec-bankruptcies	Number of public record bankruptcies
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.
tax-liens	Number of tax liens

Table 1: Description of Numerical Variables

Numerical Variables	Description
tot-coll-amt	Total collection amounts ever owed
tot-cur-bal	Total current balance of all accounts
tot-hi-cred-lim	Total high credit/credit limit
total-acc	The total number of credit lines currently in the borrower's credit file
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
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-int	Interest received to date
total-rev-hi-lim	Total revolving high credit/credit limit
verification-status	Indicates if income was verified by LC, not verified, or if the income source was verified

Categorical Variables:

Table 2: Description of Categorical Variables

Categorical Variables	Description
addr-state	The state provided by the borrower in the loan application
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.
initial-list-status	The initial listing status of the loan. Possible values are W, F
loan-status	Current status of the loan
purpose	A category provided by the borrower for the loan request.
sub-grade	LC assigned loan subgrade
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
verification-status	Indicates if income was verified by LC, not verified, or if the income source was verified

Date Variables:

Table 3: Description of Date Variables

Date Variables	Description
earliest-cr-line	The month the borrower's earliest reported credit line was opened
issue-d	The month which the loan was funded
last-credit-pull-d	The most recent month LC pulled credit for this loan
last-pymnt-d	Last month payment was received

1.2 Defining Default

Table 4 shows the total observations of different loan status. Lending club provides description on different loan status (see What-do-the-different-Note-statuses-mean-).

Table 4: Counting the factors of loan status

Loan-status	Count
Charged Off	150896
Current	20339
Default	81
Fully Paid	672620
In Grace Period	350
Late (16-30 days)	106
Late (31-120 days)	513

- **Charged Off:** Loan for which there is no longer a reasonable expectation of further payments. Upon Charge Off, the remaining principal balance of the Note is deducted from the account balance.
- **Current:** Loan is up to date on all outstanding payments.
- **Default:** Loan has not been current for an extended period of time.
- **Fully paid:** Loan has been fully repaid, either at the expiration of the 3 or 5 year term or as a result of a prepayment.
- **In Grace Period:** Loan is past due but within the 15-day grace period.
- **Late (16-30):** Loan has not been current for 16 to 30 days.

- **Late (31-120):** Loan has not been current for 31 to 120 days.

  According to *Lending Club* documentation, when borrowers miss a loan payment, their loan will move from "current" to "late" status. After borrowers miss several payments, the loan will enter "default" status and, when there is no longer a reasonable expectation of further borrower payments, the loan will be "Charged-off". Charge Off typically occurs when a loan is 120 days or more past due and there is no reasonable expectation of sufficient payment to prevent the charge off.

In other terms, The loan corresponds to a "Current" status which is still being reimbursed in a timely manner. A payment which is between 16 and 120 days overdue refers to a "Late" status. The loan is considered to be in "Default", if the payment is delayed by more than 121 days. If *Lending Club* has decided that the loan will not be paid off, then it is given the status of "Charged-Off".

From the above information, as both of the status "Late (31-120)" and "Default" include delay of payment for 120 days, I will labeled them as "Charged-off" in my analysis. Also, these time-spans imply that 4 months after the term of each loan has ended, every loan ends in two states-"fully paid" or "Charged-Off". I will consider the loans that have expired at the time of analysis. For example, the status of loan funded in December 2015 with 60 months maturity, cannot be known until December 2020. Therefore, to simplify the problem in my analysis, I will include the loans that have been funded for 36 months term.

2 Data Appendix

2.1 Data Wrangling

Deleted variables: First, I have replaced the blank cells of the data frame with the "NA" values. Then, I have removed those variables which does not have any other values rather

than NA. In addition to that, I have removed several variables for different reasons which are given in the following:

- pymnt-plan: Payment plan variable has same categorical value for all the observation. Since it has the same categorical value for all the observation, the variable is not going to give us any significant information about the borrowers. therefore, I have decided not to include this in my data frame.
- policy-code : Same numerical value for all observation.
- application-type: 844,394 borrowers have individual application type. The joint application type is insignificant compared to individual. Since we have almost same categorical value for all the observation, the inclusion of this variable will not give us any significant insight.
- url: The url variable does not have any relevant information regarding the default. As a result, I have deleted the url variable.
- desc: Description variable records the borrowers' description for the loan purpose. There are other variable which also records the borrower's loan purpose in a short term. From my analysis, I have noticed that the loan purpose variable encrypts the essence of the loan description. Hence, I have selected loan-purpose variable to include as a features, and delete the description variable.
- zip-code: There are two variables related to the location of the borrowers; zip code and state. I have used the state as the location of the borrowers and removed the zip-code variable.
- title and emp-title: The emp-title variable records the employment description of the borrowers. The loan title provided by the borrower is registered in the title variable. These two variable have large share of unique categorical values. That's why, I have deleted these variable.

- grade: Lending Club assigns loan grade from A to G for the borrowers. Sub-grade is classified into 5 sub-group for each of the grade. For example, grade A is divided into A1 to A5 sub-group. Overall, there are 35 subgroup. Since both variables are related to each other, I have excluded the "group" variable from the model.

- recoveries and collection-recovery-fee: The recoveries and collection-recovery-fee variables document post charge off gross recovery. Both of the variables have almost same values for each observation.

- Variables with similar observation: I have excluded "collections-12-mths-ex-med","out-prncp-inv", "hardship-flag", "out-prncp" variables for having almost identical value for each observation.

More than 60 percent missing values: There are 31 variables with more than 60 percent missing values. I have removed 28 of them. The variable "mths-since-last-major-derog" keeps record of months since most recent 90-day or worse rating, and the variable "mths-since-last-record" registers the number of months since last public record. The NA values might indicate that there was no record of missed payment. Therefore, removing these two variables may cause loss of valuable information. I have replaced the missing values of these two variables with 0.

Table 5: Variables More than 60 Percent Missing Values

variable name	Count of Missing Values
mths-since-last-record	711494
mths-since-last-major-derog	623184
annual-inc-joint	844396
dti-joint	844398
e open-acc-6m	823535
open-act-il	823535
open-il-12m	823535
open-il-24m	823535
mths-since-rcnt-il	824097
total-bal-il	823535
il-util	826290
open-rv-12m	823535
open-rv-24m	823535
max-bal-bc	823535
all-util	823535
inq-fi	823535
total-cu-tl	823535
inq-last-12m	823535
mths-since-recent-bc-dlq	637275
mths-since-recent-revol-delinq	553957
deferral-term	840158
hardship-amount	838931
hardship-length	840158
hardship-dpd	840158
orig-projected-additional-accrued-interest	839545
hardship-payoff-balance-amount	838931
hardship-last-payment-amount	838931
settlement-date	825230
settlement-percentage	825230
hardship-type	840159
hardship-reason	840158
hardship-status	840158
hardship-start-date	840158
hardship-end-date	840158
hardship-loan-status	840161
settlement-status	825230
settlement-term	825230
verification-status-joint	844396
next-pymnt-date	823516

2.2 Data Transformation

Character to Numeric Variables: The variables int-rate and revol-util are were recorded as character variables. I have changed them to numeric variables by using "dplyr" package in R.

Character to Date Variables: All the date variables were given in string "yearmonth" format. I have transformed them into date format.

2.3 Handling the Missing Values

There are 61 variables with less than 60 percent missing observations.

Replacing Missing Values of Numerical Variables: I have replaced the missing values of the numerical variables in two ways; replace with median, and replace with zero.

- Replacing missing values with median:
  - ☐ funded-amnt
  - ☐ loan-amnt
  - ☐ funded-amnt-inv
  - ☐ annual-inc
  - ☐ dti
  - ☐ fico-range-low
  - ☐ tot-hi-cred-lim

- Replacing missing values with zero:
  - ☐ mths-since-last-delinq
  - ☐ mths-since-last-record
  - ☐ mths-since-last-major-derog
  - ☐ installment
  - ☐ inq-last-6mths
  - ☐ open-acc
  - ☐ pub-rec
  - ☐ revol-bal
  - ☐ total-acc
  - ☐ total-pymnt

- ☐ total-pymnt-inv
- ☐ total-rec-prncp
- ☐ total-rec-int
- ☐ last-pymnt-amnt
- ☐ last-fico-range-high
- ☐ last-fico-range-low
- ☐ acc-now-delinq
- ☐ tot-coll-amt
- ☐ acc-open-past-24mths
- ☐ bc-util
- ☐ chargeoff-within-12-mths
- ☐ delinq-amnt
- ☐ mo-sin-old-il-acct
- ☐ mo-sin-old-rev-tl-op
- ☐ mo-sin-rcnt-rev-tl-op
- ☐ mo-sin-rcnt-tl
- ☐ mort-acc
- ☐ mths-since-recent-bc
- ☐ mths-since-recent-inq
- ☐ num-accts-ever-120-pd
- ☐ num-actv-bc-tl
- ☐ num-actv-rev-tl
- ☐ num-bc-sats
- ☐ num-bc-tl
- ☐ num-il-tl
- ☐ num-op-rev-tl
- ☐ num-rev-accts
- ☐ num-rev-tl-bal-gt-0
- ☐ num-sats
- ☐ num-tl-120dpd-2m
- ☐ num-tl-30dpd
- ☐ num-tl-90g-dpd-24m
- ☐ num-tl-op-past-12m
- ☐ pct-tl-nvr-dlq
- ☐ percent-bc-gt-75
- ☐ pub-rec-bankruptcies
- ☐ tot-cur-bal
- ☐ total-rev-hi-lim
- ☐ avg-cur-bal
- ☐ bc-open-to-buy
- ☐ total-bal-ex-mort
- ☐ total-bc-limit
- ☐ total-il-high-credit-limit
- ☐ revol-util

Missing Values of Categorical Variables: The emp-length variable has 43,723 missing values. The *Lending Club* has labeled the missing observation as n/a. I will keep those observation and use the label n/a as a category. The other categorical variables have two missing values for the same observations. I have removed those two observation.

Replacing Missing Values of Date Variables: The "last-pymnt-d" and "last-credit-pull-d" have some missing observations. "Last payment date" records the date when last payment was received. "Last-credit-pull-d" registers the last date of credit inquiries. The total payment related to the missing observations of these two variables were recorded as zero. Therefore, I chose to replace those missing observations with the median.

References

Team, R. C. et al. (2013). R: A language and environment for statistical computing.