2020

Capstone Project I Milestone Report

2018 LENDING CLUB LOAN DATA

NGOC PHAN

Project Background

LendingClub (LC) is an online credit marketplace that enables a borrower to apply for a loan and an investor to select a loan to invest. When a borrower applies for a loan at LC, the company will screen the applicant. If the loan application gets approved, LC will provide the borrower the interest rate for the loan. Once the borrower accepts the loan, the loan is made available for the investors to select. The investor may choose to invest in a whole loan or a fractional part of a loan.

Problem Statement

This project studies the 2018 LC loan data to assist prospective investors in the decision-making process by:

- Exploring the characteristics of fully paid and default loans.
- Estimating the return on investment (ROI) and loss of investment.
- Developing machine learning algorithms that predicts the likeliness of a default loan as well as ROI.

Dataset

The loan datasets are collected on LendingClub website at https://www.lendingclub.com/info/download-data.action that include 2018 loan data by quarter. Each dataset is in csv format and stored in a zip file. There are 495,242 rows and 144 columns in the dataset. All columns contain cleaned data, and no duplicates have been found. The following rows are ignored when reading the csv files:

- The first row of each csv file that contains general note.
- The last two rows of each csv file that contain the total amount funded in policy code 1 and 2.

Columns id, member_id, url, and desc contain no values and are excluded from the data.

Data Preparation

Jupyter notebook for data wrangling can be found at https://github.com/nphan20181/Loan-Default-Prediction/blob/master/loan_data_wrangling.ipynb.

Removing Columns

Remove columns that contain information that is not useful or readily available at the time a loan is issued. Examples of those columns are loan id, hardship flag, total received interest, etc. There are 98 columns that are subjects of interest such as loan amount, interest rate, etc.

Removing Rows

Remove all records that do not have a loan status of fully paid, charged off or default.

Missing Values

Remove columns that have more than 25% of missing values. There are 18 columns. The missing-value percentage for those columns ranges from 54% to 96%. Since the missing-value percentage for those columns is high, it is necessary to remove the columns.

For columns that have less than or equal to 25% of missing values, fill in the median value for numerical variables and leave the values as null for non-numerical variables. There are 2 non-numerical columns and 13 numerical columns. The missing-value percentage for those columns ranges from 0.01% to 17%.

Outliers

Compute z-scores to obtain the records and variables that contain outliers (z-score < -3 or z-score > 3). There are 36,355 rows and 64 columns that contain outliers. Since the number of outliers in the dataset is high (195,492 / 495,242 = 39%), it is necessary to keep the outliers because they may contain significant information, and there are also some models that work well with outliers.

New Columns

- Loan Status Flag

  Categorize loan status into 2 categories:

  - Fully paid

    - Default (Default, Charged Off)

- Return on Investment (ROI) = total payment amount – loan amount
- Months in Loan = last payment date – loan issued date

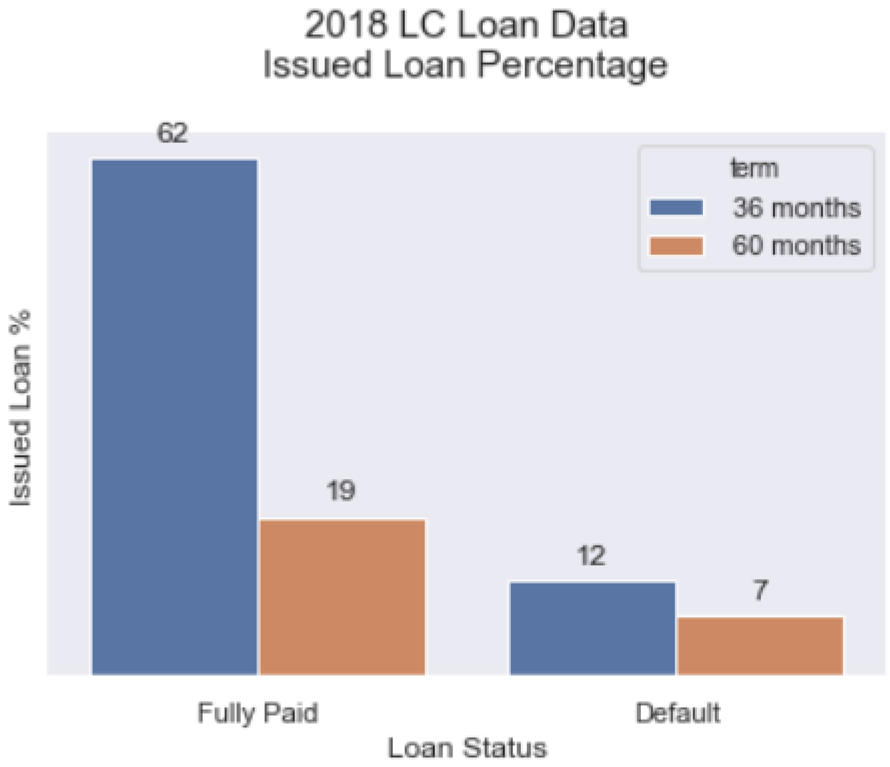
Exploratory Data Analysis

Jupyter notebook for data analysis can be found below:

- Descriptive statistics: https://github.com/nphan20181/Loan-Default-Prediction/blob/master/lc_loan_data_story.ipynb
- Inferential statistics: https://github.com/nphan20181/Loan-Default-Prediction/blob/master/lc_inferential_stats.ipynb

Issued Loans

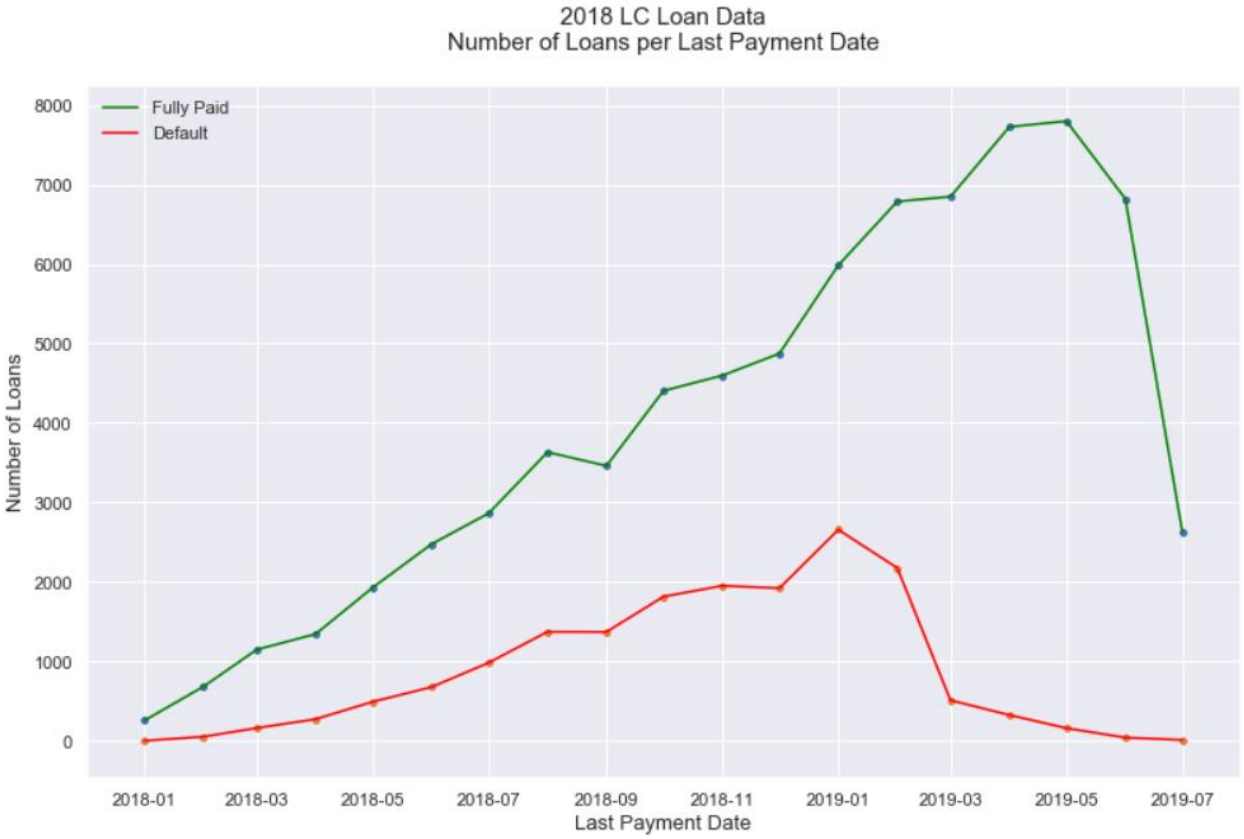
In 2018, the total number of fully paid and default loans were 93,853 loans in which 81% of the loans were paid-off and 19% of the loans were defaulted. 74% of the borrowers were on 36 months loan term while 26% of the borrowers were on 60 months loan term.



Loan’s Last Payment Date

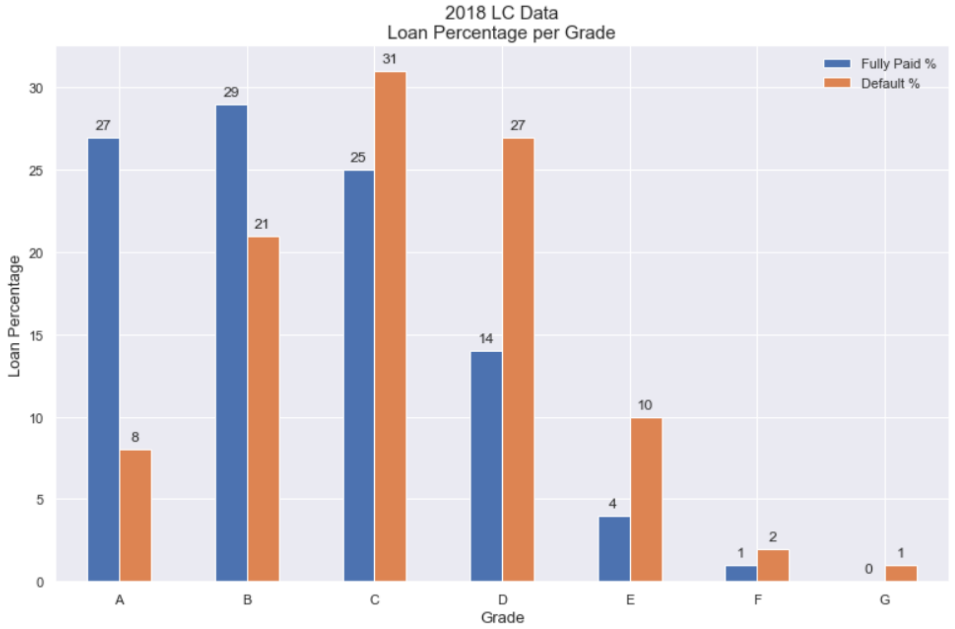
All fully paid loans are paid off with 17 months. The trend is upward from January 2018 to May 2019 and is downward after May 2019.

For default loans, the trend line is upward from January 2018 to January 2019 and is downward after January 2019. More borrowers made their last payment around the end of 2018 and the first two months of 2019 before defaulting on their loans.

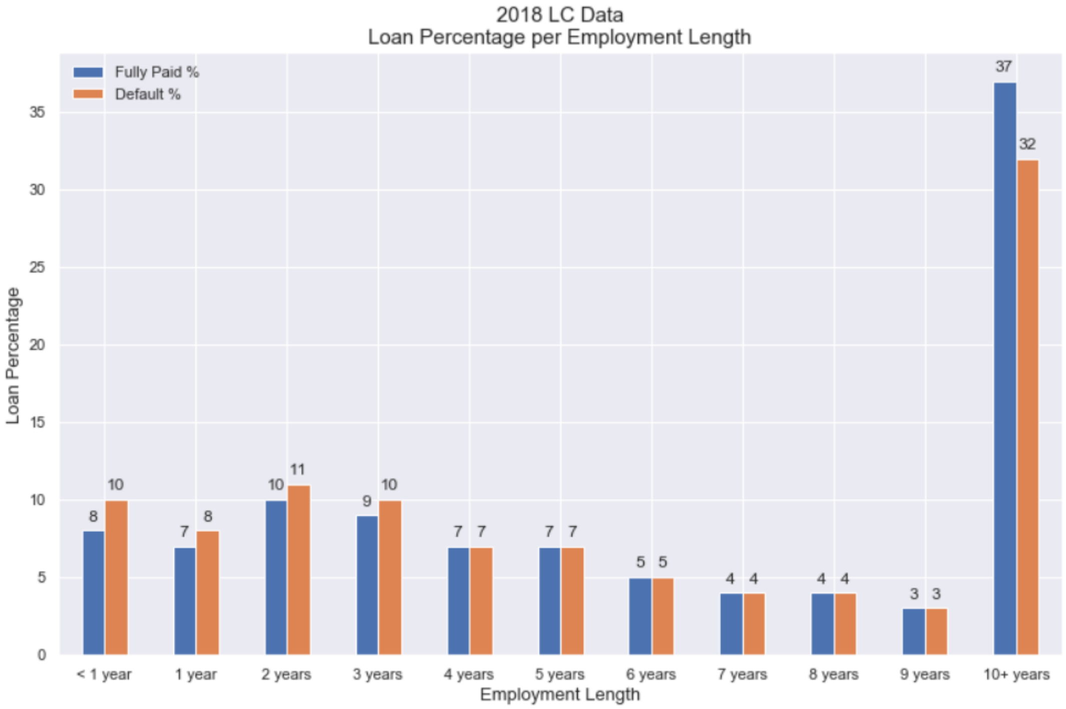


Data Distribution

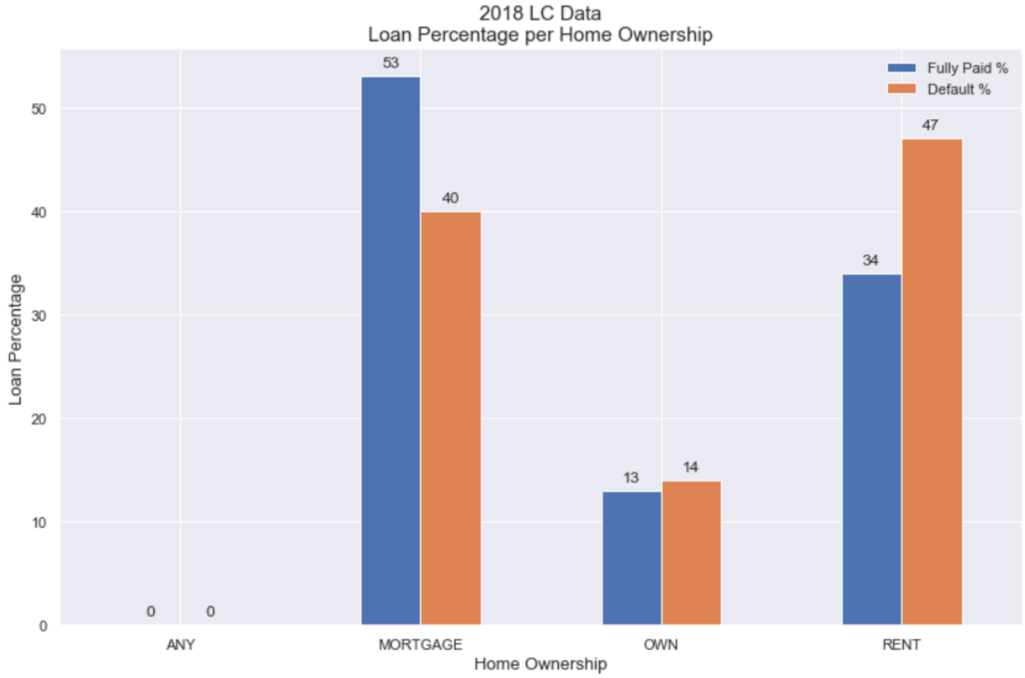
Most fully paid loans have a grade of A, B, or C while most default loans have a grade of B, C, D.



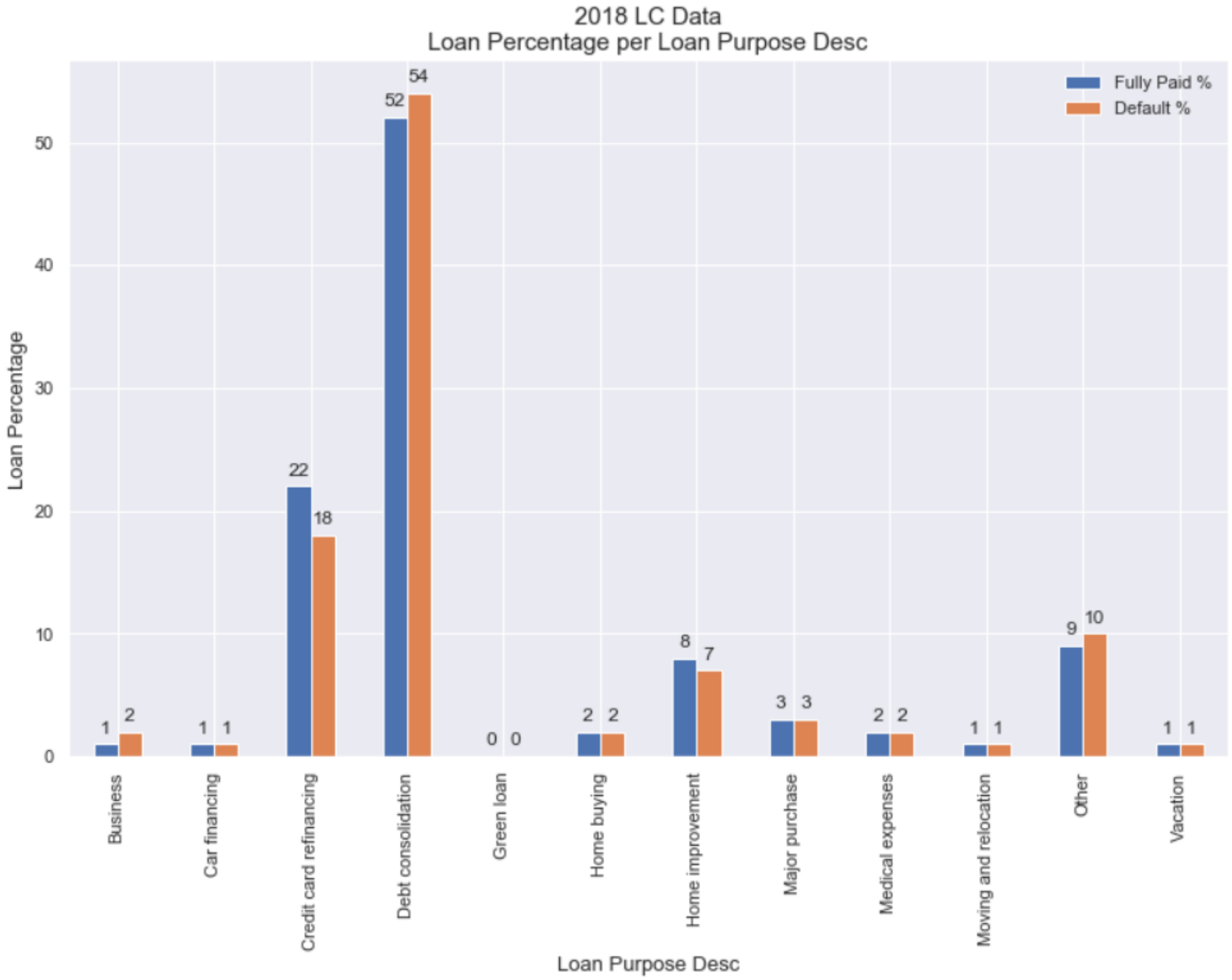
The distribution of employment length for both fully paid and default loans are similar. Both have high percentage of borrowers who have been employed 10 or more years.



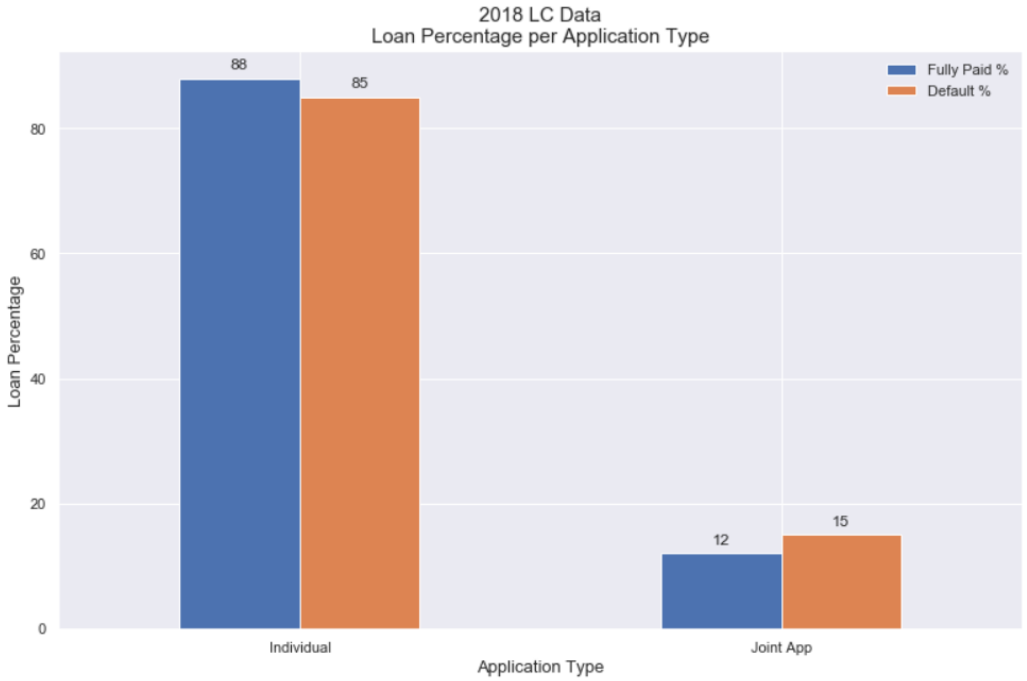
Most fully paid loans have borrowers who are on mortgage while most default loans have borrowers who are on house rental.



The distribution of loan purpose for both fully paid and default loans are similar. Many loans were issued to borrowers for debt consolidation and credit card refinancing.

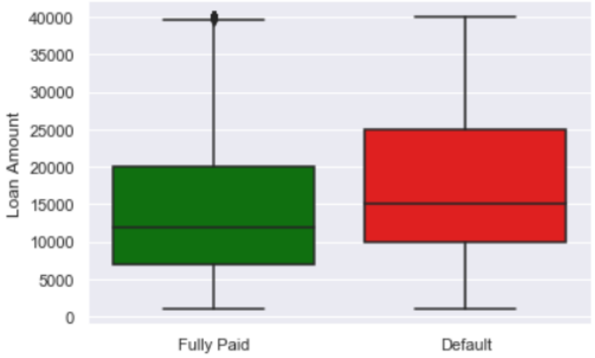


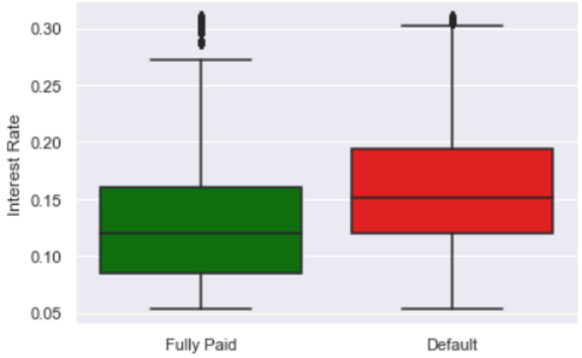
Likewise, the distribution of loan's application type are similar for both default and fully paid loans. There are more individual applications than joint applications.

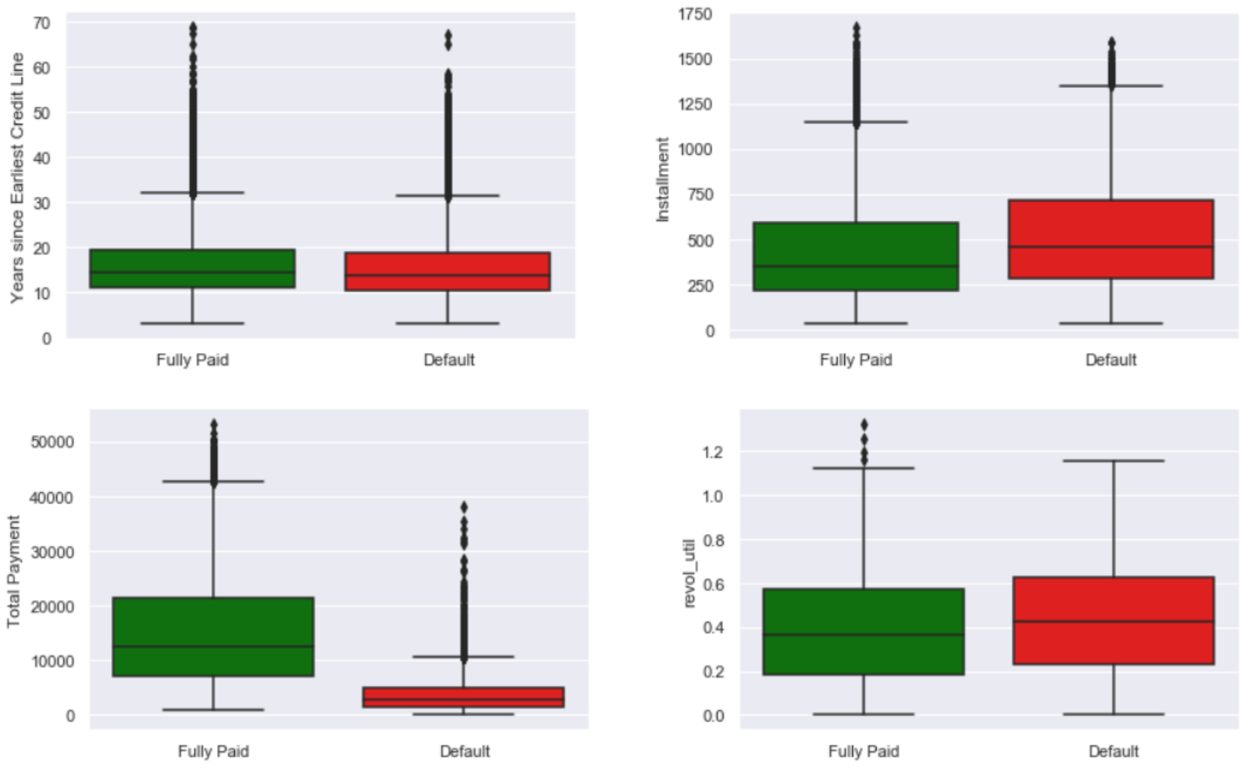


Data Location

Most fully paid loans have lower loan amount, interest rate, and installment compared to those of default loans. The total payment for most fully paid loans are also higher than that of default loans. Years since earliest credit line and the amount of credit the borrower is using relative to all available revolving credit (revol_util) are almost the same for both fully paid and default loans.





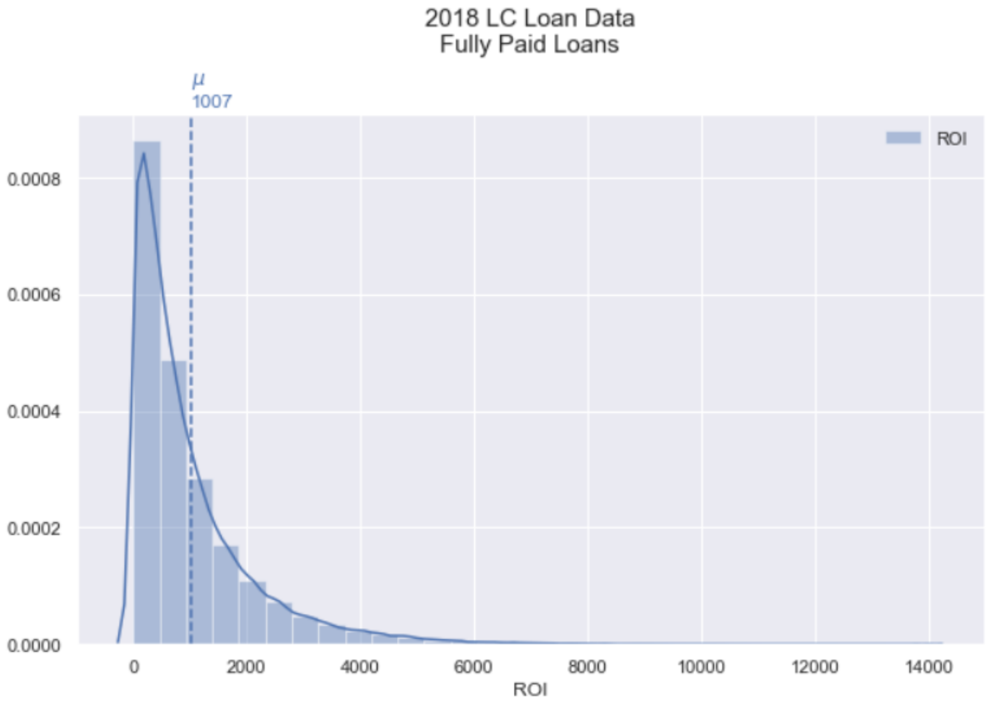


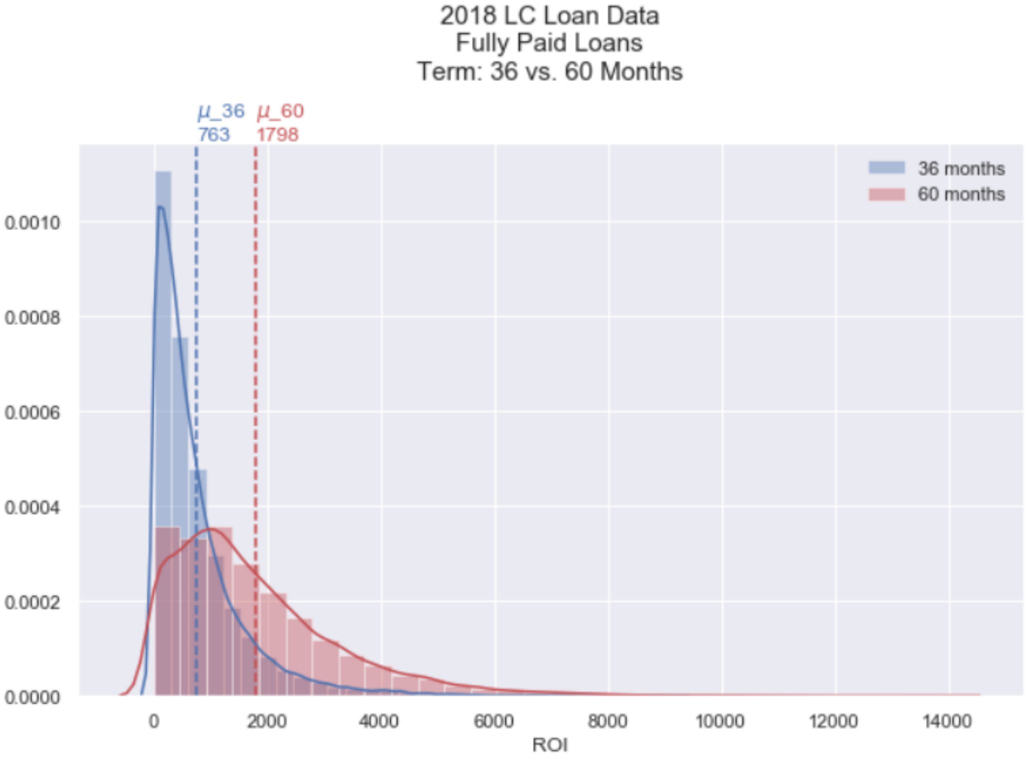
Return on Investment (ROI)

For fully paid loans, we are 95% confident that the average ROI is:

- between 990 and 1015 dollars, regardless of loan term.
- between 756 and 770 dollars for 36 months loan.
- between 1776 and 1820 dollars for 60 months loan.

The t-test result, (-116.97, 0.0), also indicates that there is a significant difference on mean ROI between 36 and 60 months loan term.





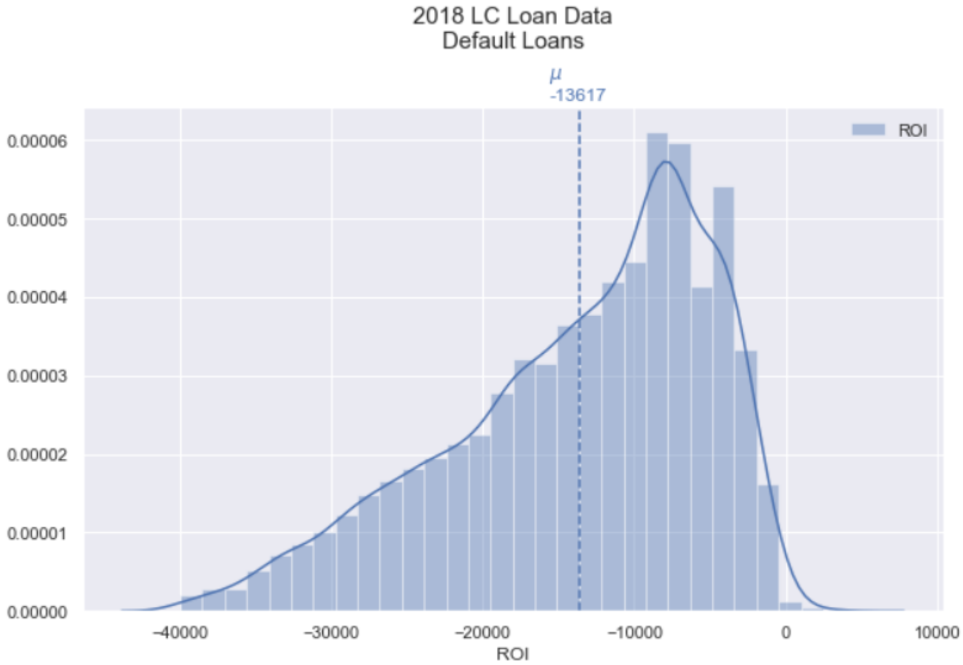
Loss of Investment

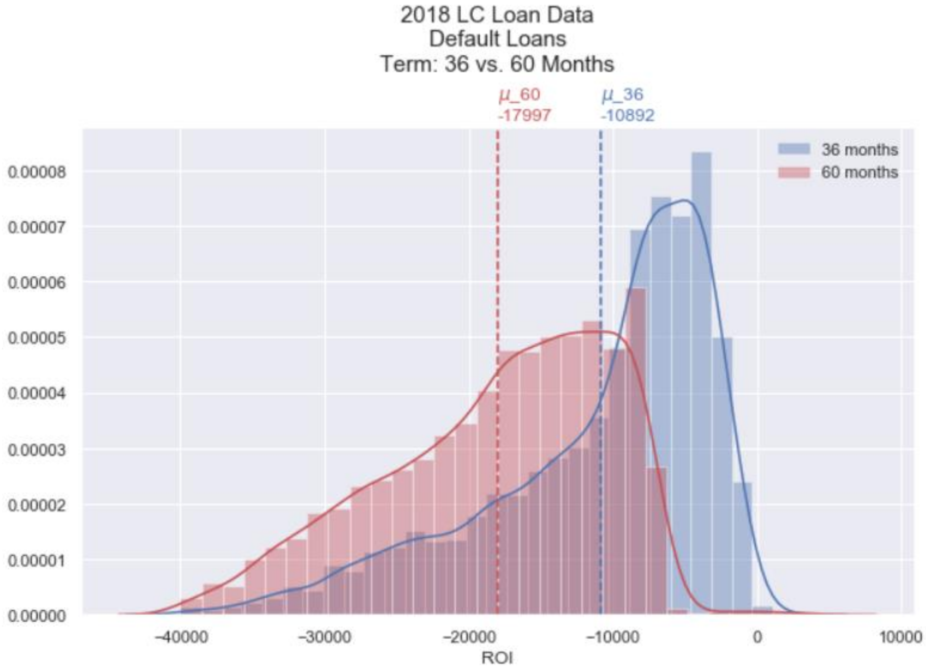
For default loans, we are 95% confident that the average loss is:

- between 13,490 and 13,744 dollars, regardless of loan term.

- between 10,744 and 11,040 dollars for 36 months loan.
- between 17,810 and 18,185 dollars for 60 months loan.

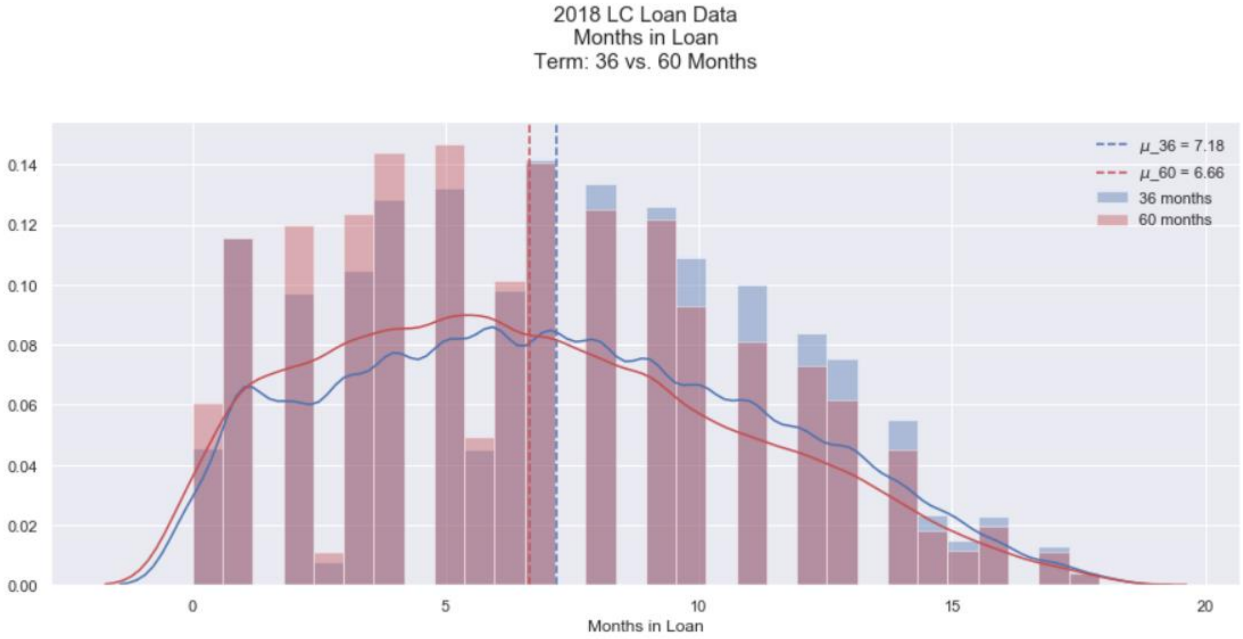
The t-test result, (58.17, 0.0), also indicates that there is a significant difference on average loss between 36 months and 60 months loan term.





Months in Loan

Both 36 months and 60 months loan terms have an average months in loan close to 7 months. All borrowers of both fully paid and default loans stayed in the loan less than 20 months which is shorter than their loan term. The t-test result, (16.67, 2.85e-62), also indicates that there is a significant difference on the mean months in loan for both 36 and 60 months loan terms.



Investment Portfolio: Estimating Loss of Investment

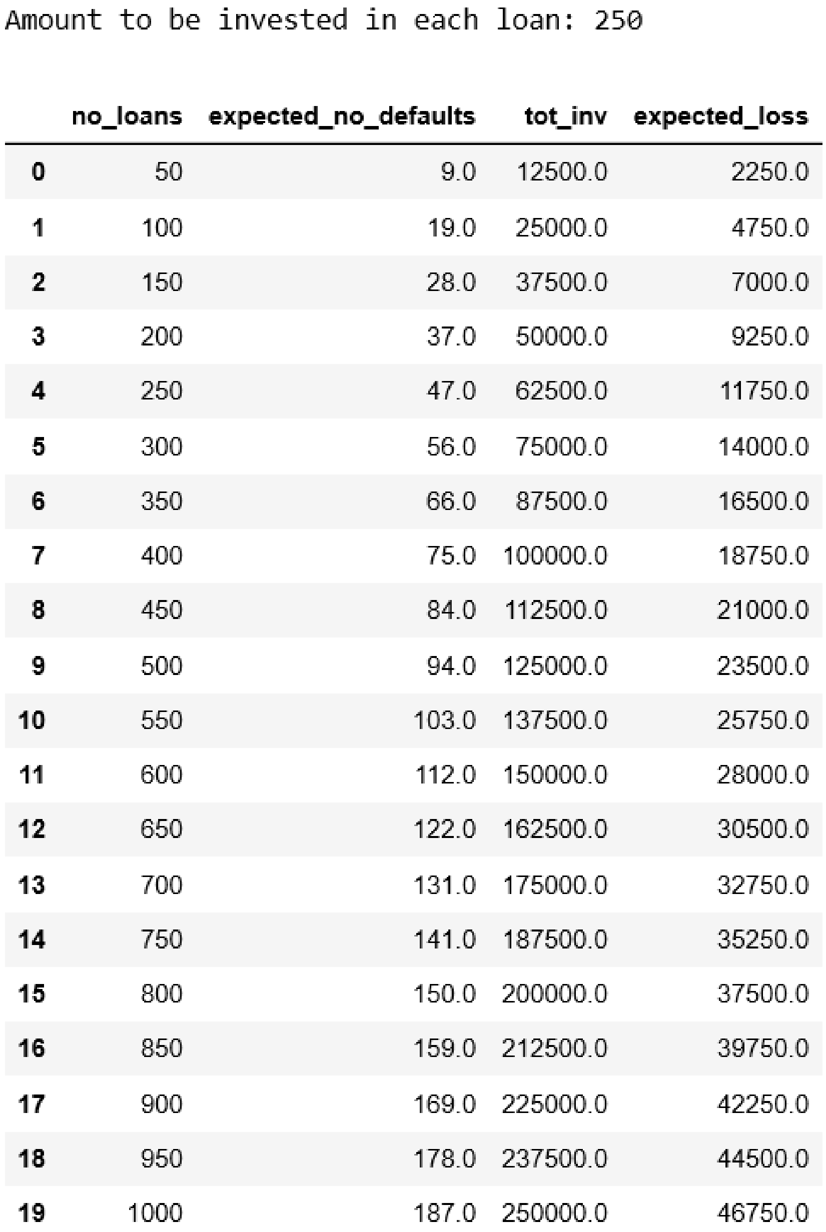
Apply binomial distribution to estimate the number of default loans which is the mean of the distribution for a given sample. The table below shows the expected number of defaults and loss based on the number of invested loans. The data generated is based on the following assumptions:

- The lender invests $250 in each loan.
- The lender loses all money invested in a loan if the loan is default.
- There is no other hidden fees or costs.

Variables

- ***no_loans***: number of loans to be invested in

- ***expected_no_defaults***: the expected number of defaults is the mean of the binomial distribution with the probability of default of 17,589 / 93,853.
- ***tot_inv***: total amount invested = amount invested in each loan * no_loans
- ***expected_loss***: expected loss = amount invested in each loan * expected_no_defaults



```
Amount to be invested in each loan: 250
```

	no_loans	expected_no_defaults	tot_inv	expected_loss
0	50	9.0	12500.0	2250.0
1	100	19.0	25000.0	4750.0
2	150	28.0	37500.0	7000.0
3	200	37.0	50000.0	9250.0
4	250	47.0	62500.0	11750.0
5	300	56.0	75000.0	14000.0
6	350	66.0	87500.0	16500.0
7	400	75.0	100000.0	18750.0
8	450	84.0	112500.0	21000.0
9	500	94.0	125000.0	23500.0
10	550	103.0	137500.0	25750.0
11	600	112.0	150000.0	28000.0
12	650	122.0	162500.0	30500.0
13	700	131.0	175000.0	32750.0
14	750	141.0	187500.0	35250.0
15	800	150.0	200000.0	37500.0
16	850	159.0	212500.0	39750.0
17	900	169.0	225000.0	42250.0
18	950	178.0	237500.0	44500.0
19	1000	187.0	250000.0	46750.0

Appendix

Appendix A – Variable Description

Column Name	Description
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
acc_open_past_24mths	Number of trades opened in past 24 months.
all_util	Balance to credit limit on all trades
annual_inc	The self-reported annual income provided by the borrower during registration.
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
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
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
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.
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.
earliest_cr_line	The month the borrower's earliest reported credit line was opened
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.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. LC values are: RENT, OWN, MORTGAGE, OTHER
il_util	Ratio of total current balance to high credit/credit limit on all install acct
inq_fi	Number of personal finance inquiries
inq_last_12m	Number of credit inquiries in past 12 months
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
issue_d	The month which the loan was funded

last_pymnt_d	Last month payment was received
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.
loan_status	Current status of the loan
max_bal_bc	Maximum current balance owed on all revolving accounts
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_rcnt_il	Months since most recent installment accounts opened
mths_since_recent_bc	Months since most recent bankcard account opened.
mths_since_recent_inq	Months since most recent inquiry.
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_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
open_acc	The number of open credit lines in the borrower's credit file.
open_acc_6m	Number of open trades in last 6 months
open_il_12m	Number of installment accounts opened in past 12 months
open_il_24m	Number of installment accounts opened in past 24 months
open_act_il	Number of currently active installment trades
open_rv_12m	Number of revolving trades opened in past 12 months
open_rv_24m	Number of revolving trades opened in past 24 months
out_prncp	Remaining outstanding principal for total amount funded
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
pymnt_plan	Indicates if a payment plan has been put in place for the loan
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
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
title	The loan title provided by the borrower
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_bal_il	Total current balance of all installment accounts
total_bc_limit	Total bankcard high credit/credit limit
total_cu_tl	Number of finance trades
total_il_high_credit_limit	Total installment high credit/credit limit
total_pymnt	Payments received to date for total amount funded