2020

Capstone Project I
Final 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 LendingClub investors in the loan selection process by:

- Exploring the characteristics of fully paid and default loans.
- Exploring return on investment (ROI) and loss of investment.
- Estimating the number of defaults for a given sample.
- Developing machine learning algorithms that classifies a loan as default or fully-paid and estimates the length of a loan.

Dataset

The loan datasets are collected on LendingClub website 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

Removing Columns

Remove columns that contain information that is not useful or not readily available at the time a loan is issued. Examples of those columns are loan id, hardship flag, total received interest, 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.

New Columns

- Loan Status Flag

  Categorize loan status into 2 categories:

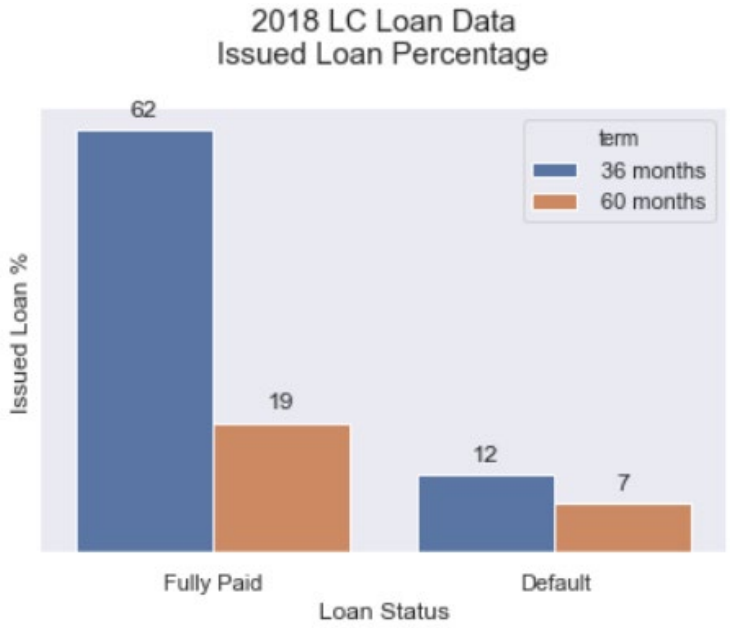
  - Fully paid
  - Default (Default, Charged Off)
- Return on Investment (ROI) = total payment amount – loan amount

- Months in Loan = last payment date – loan issued date

Exploratory Data Analysis

Issued Loans

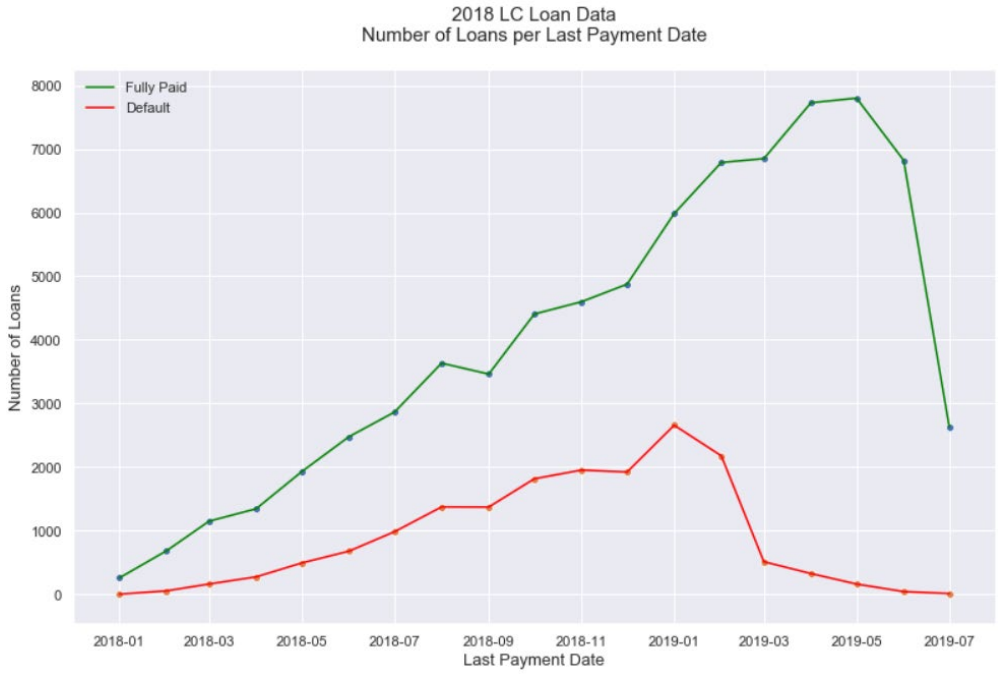
In 2018, the total number of fully paid and default loans were 93,853 loans in which 81% of the loans were paid-off and 19% of the loans were defaulted. 74% of the borrowers were on 36 months loan term while 26% of the borrowers were on 60 months loan term.



Loan's Last Payment Date

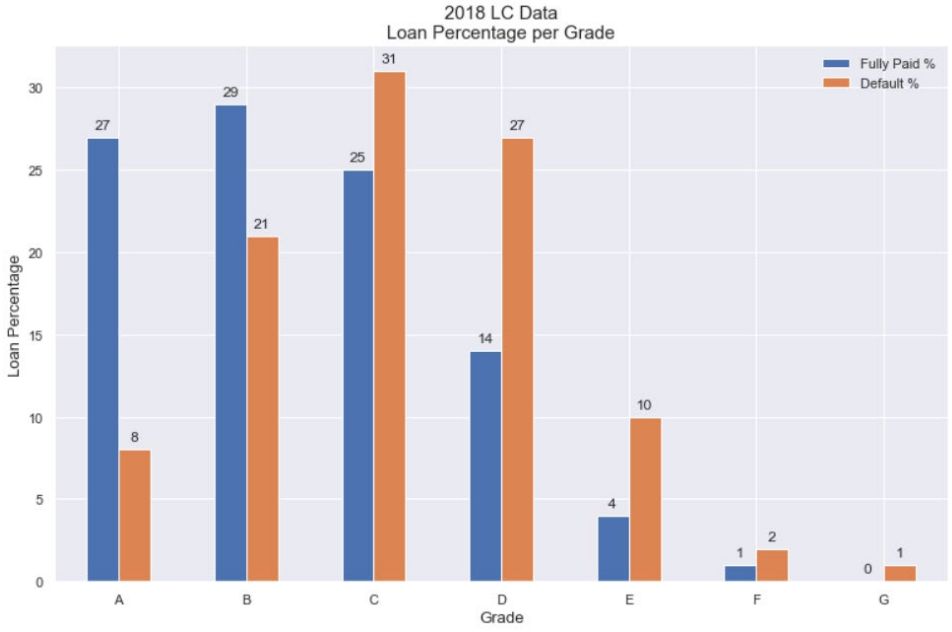
All fully paid loans are paid off within 17 months. The trend is upward from January 2018 to May 2019 and is downward after May 2019.

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.

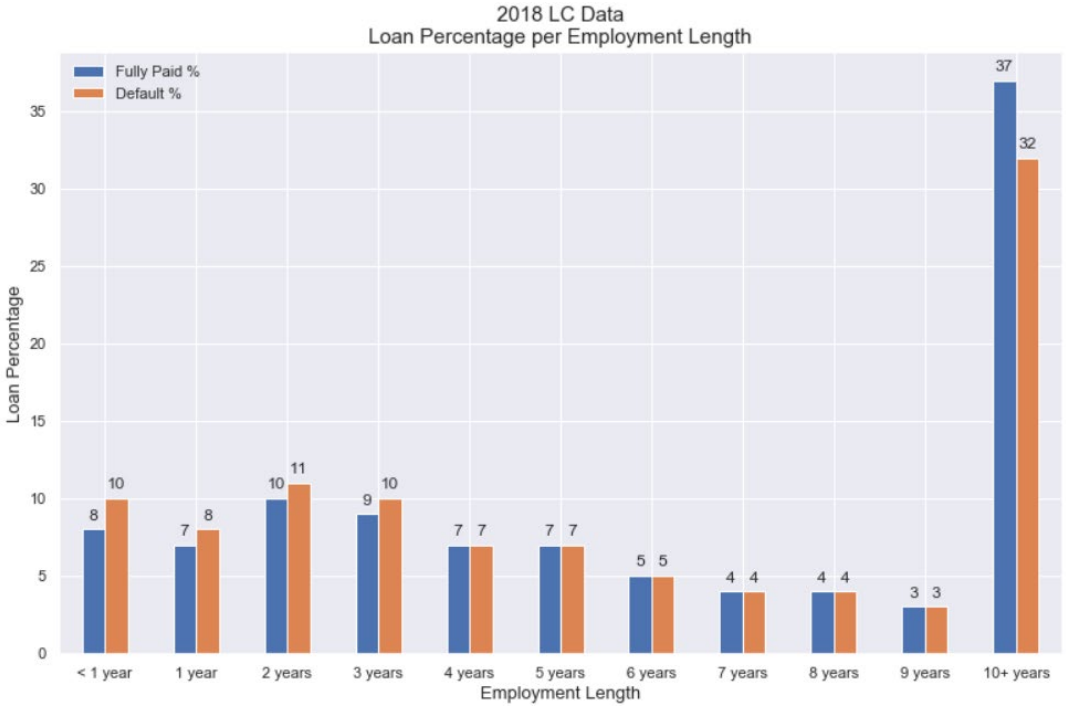


Data Distribution

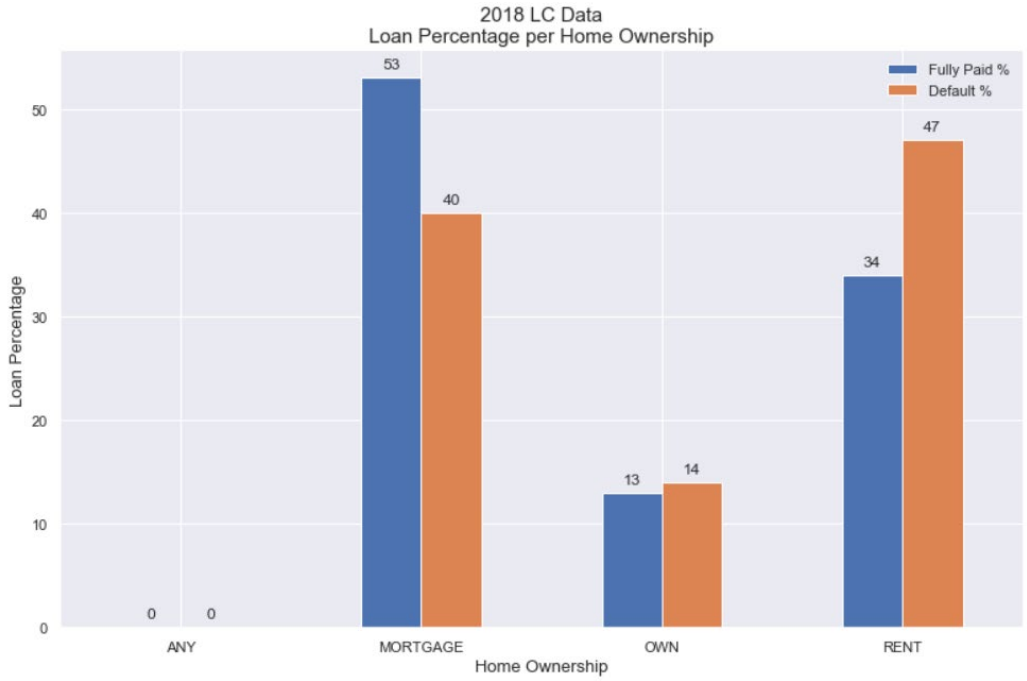
Most fully paid loans have a grade of A, B, or C while most default loans have a grade of B, C, or D.



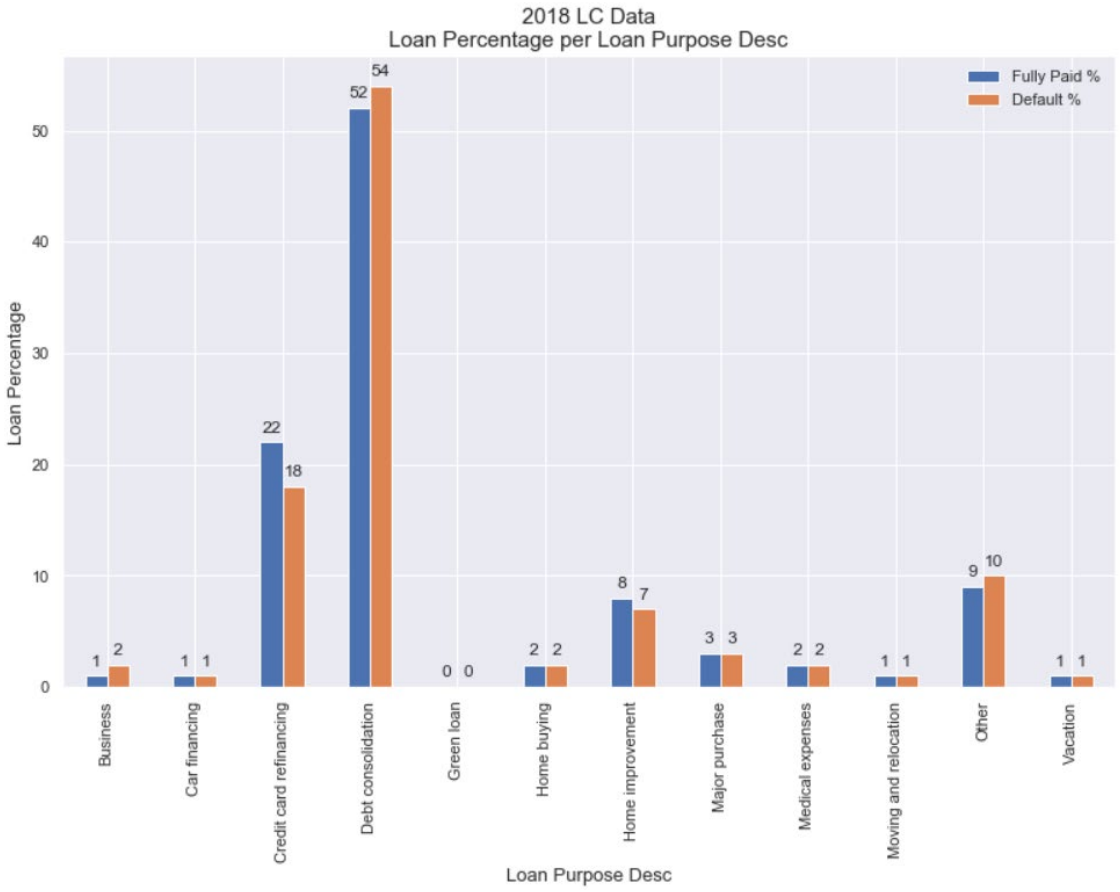
The distribution of employment length for both fully paid and default loans are similar. Both have high percentage of borrowers who have employed for 10 or more years.



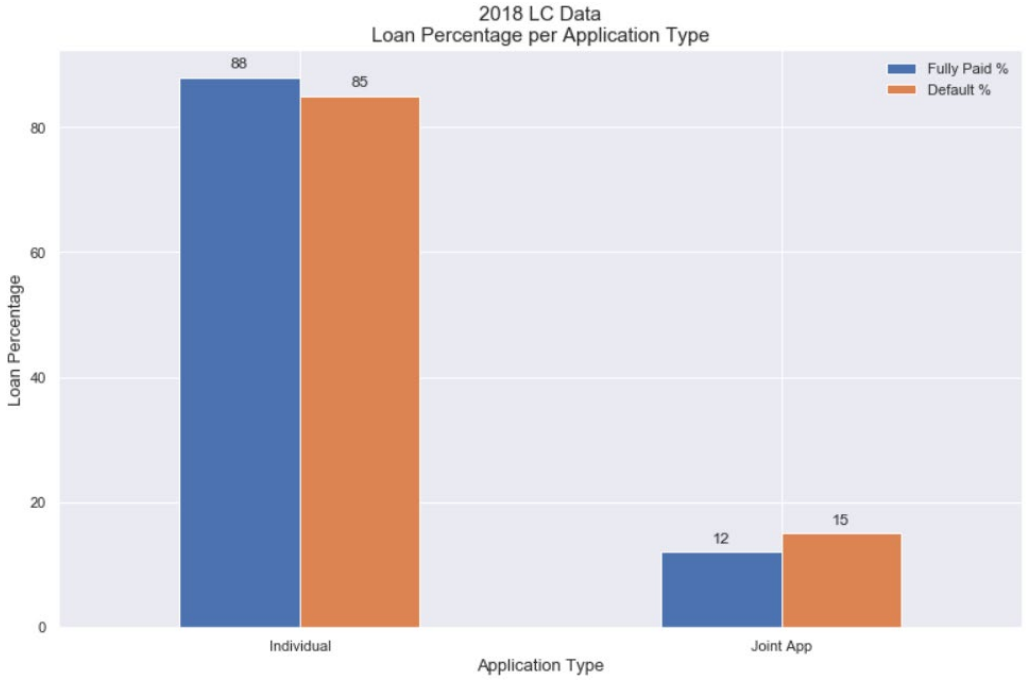
Most fully paid loans have borrowers who are on mortgage while most default loans have borrowers who are on house rental.



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.

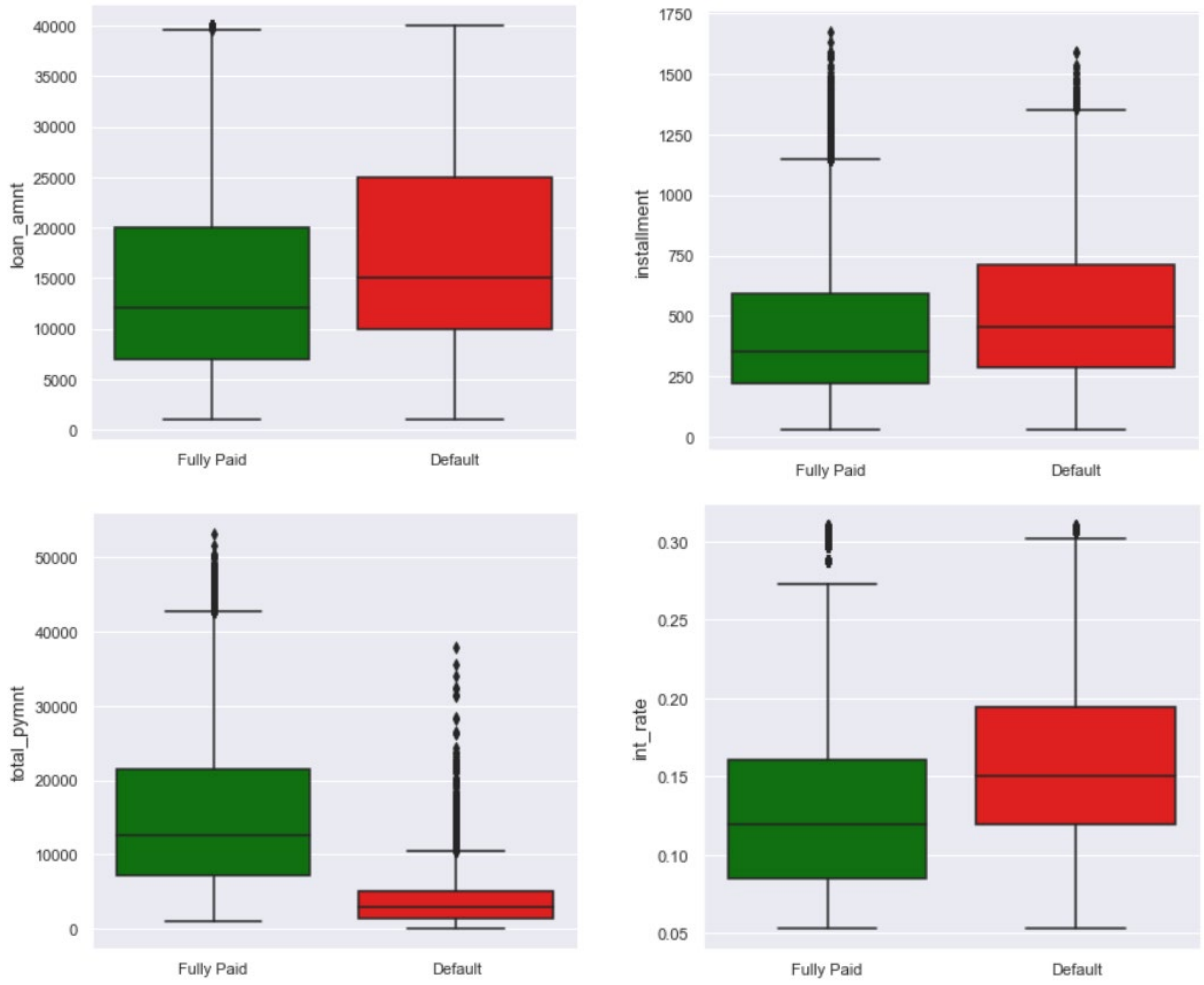


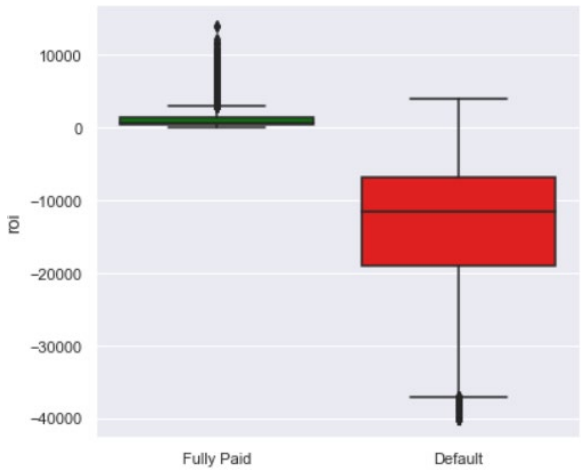
Likewise, the distribution of loan's application type are similar for both default and fully paid loans. There are more individual applications than joint applications.

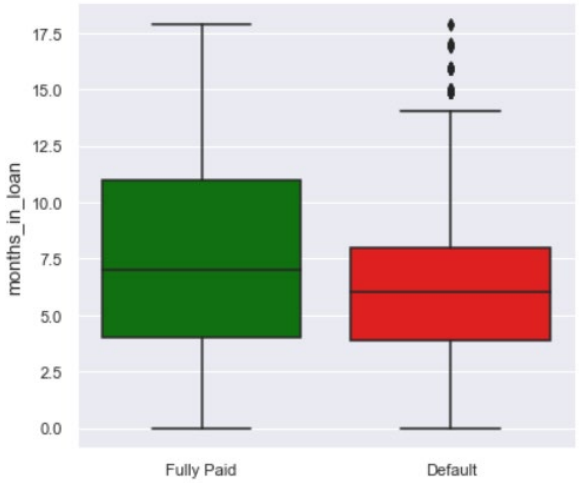


Data Location

Most fully paid loans have lower loan amount, interest rate, and installment compared to those of default loans. The total payment and return on investment (ROI) for most fully paid loans are also higher than that of default loans. Though months-in-loan for both fully paid and default loans seems to overlap, most fully paid loans tend to have longer months-in-loan.





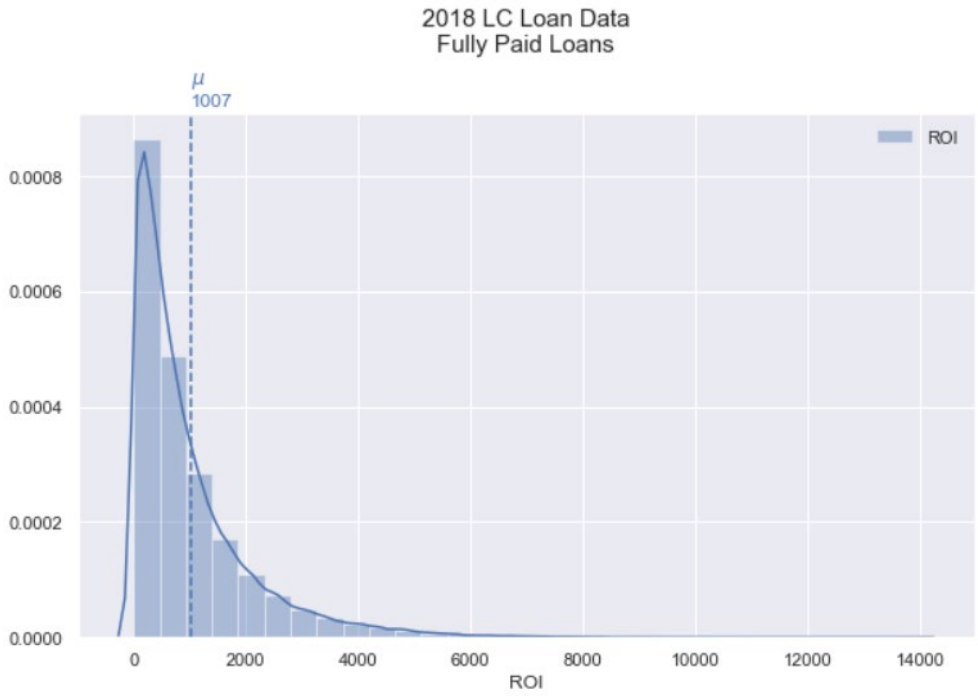


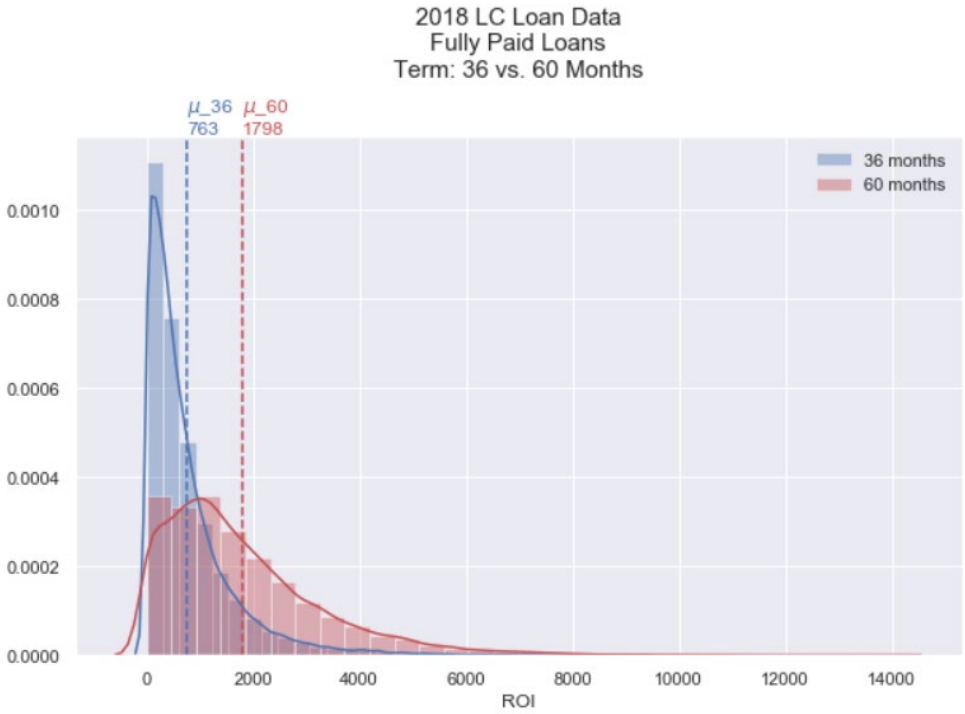
Return on Investment (ROI)

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.



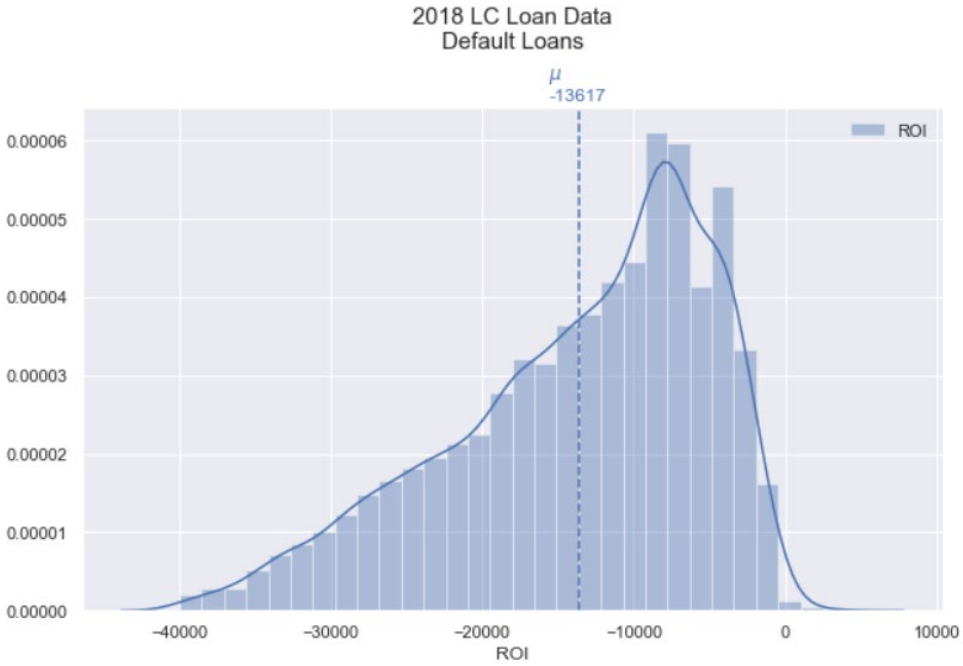


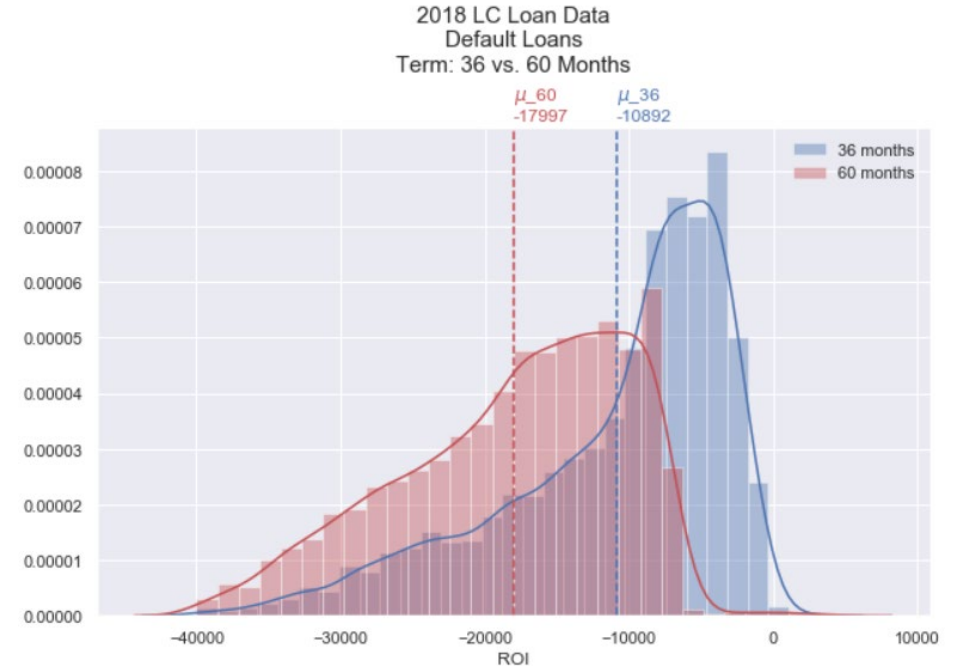
Loss of Investment

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.

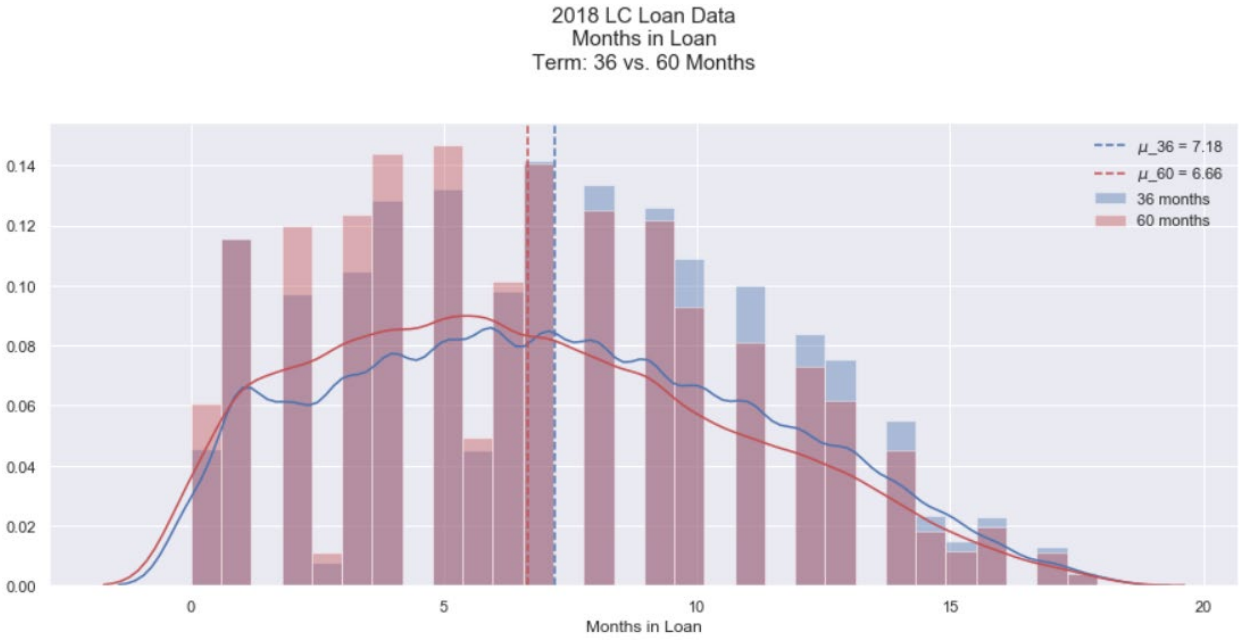




Months in Loan

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.



Estimating Number of Defaults for a Given Sample

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 for a given sample.

no_loans	expected_no_defaults
5	1.0
10	2.0
15	3.0
20	4.0
25	5.0
30	6.0
35	7.0
40	7.0

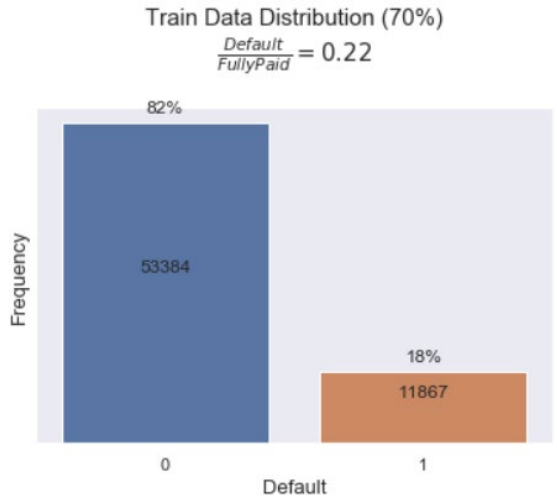
where

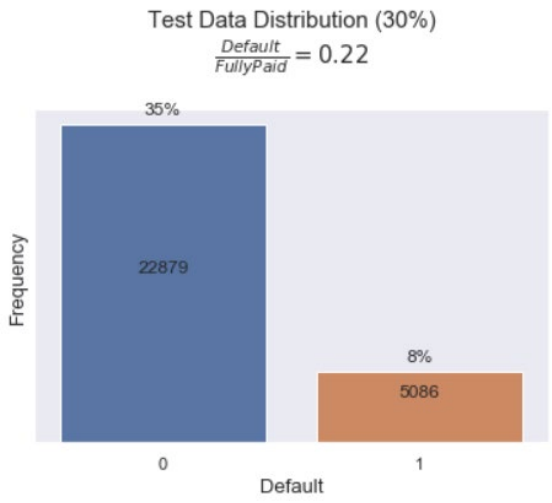
- ***no_loans***: number of loans to be invested
- ***expected_no_defaults***: the expected number of defaults is the mean of the binomial distribution with the probability of default of 18.74%.

Machine Learning Models

Train vs. Test Dataset

Apply stratified sampling to split dataset into 70% training and 30% testing. The distribution of train and test data are illustrated below:





Model Development

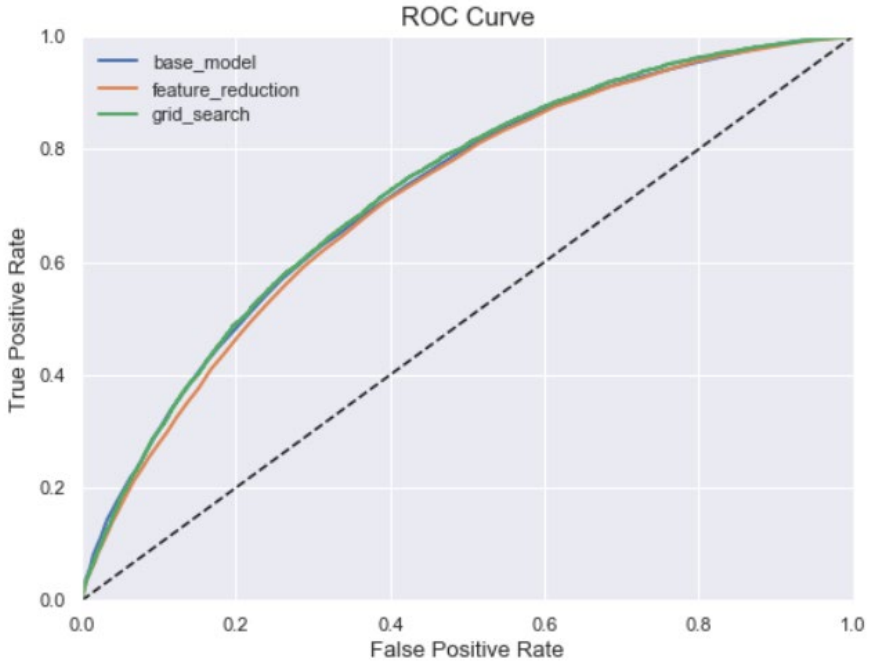
Develop the following three models for both Random Forest classifier and regressor:

Base Model	Model 1	Model 2
• Use all features • Use the following parameter: param = {'bootstrap': [True], 'n_estimators': [100]}	• Use same parameter as base model. • Use features with important score greater than 0.2. • Exclude some features with strong correlation.	• Use same features are model 1. • Use the best parameters from grid search.

Random Forest Classifier: Classify Loan

a. Model Evaluation & Selection

According to the plot below, the ROC curves for base model and grid search model (model 2) overlap. Model 1 (feature reduction) has lower true positive rate than that of base model and model 2.

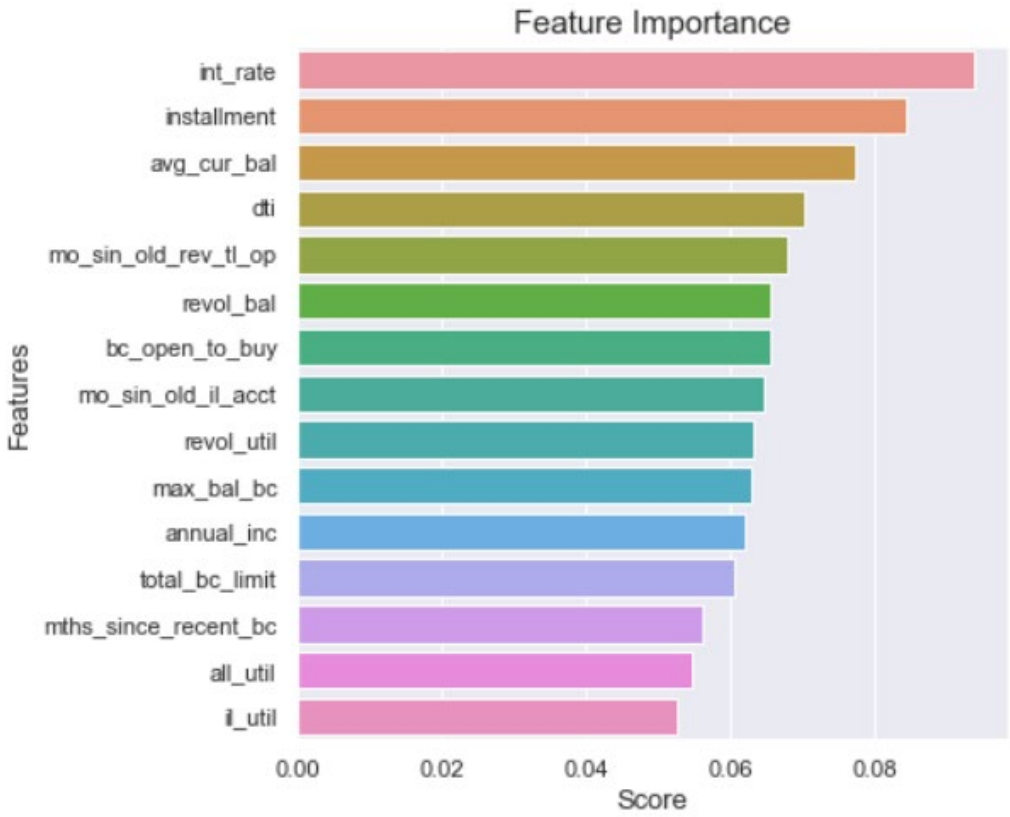


The table below indicates that the base model has highest accuracy and f1 score. However, since base model use all features in the dataset in making the prediction, it is more likely to overfit. Therefore, model 1 (feature reduction) is selected in making the prediction because the model has shortest training time and its' accuracy score and f1 score are closed to that of the base model. Model 2 (grid search) is not selected because it has lowest recall score and takes longest time to train.

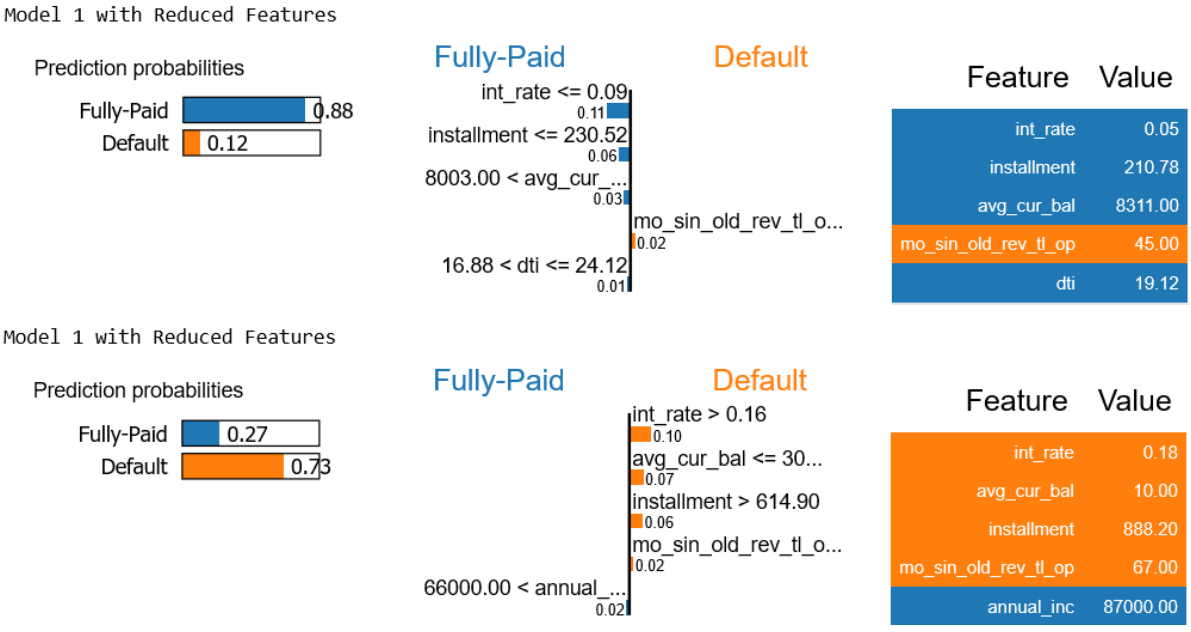
model_name	true_neg	true_pos	false_neg	false_pos	accuracy	precision	recall	f1_score	training_time
base_model	22692	241	4845	187	0.820061	0.563084	0.047385	0.087414	68
feature_reduction	22684	224	4862	195	0.819167	0.534606	0.044042	0.081381	41
grid_search	22777	145	4941	102	0.819667	0.587045	0.028510	0.054378	1179

b. Model Prediction

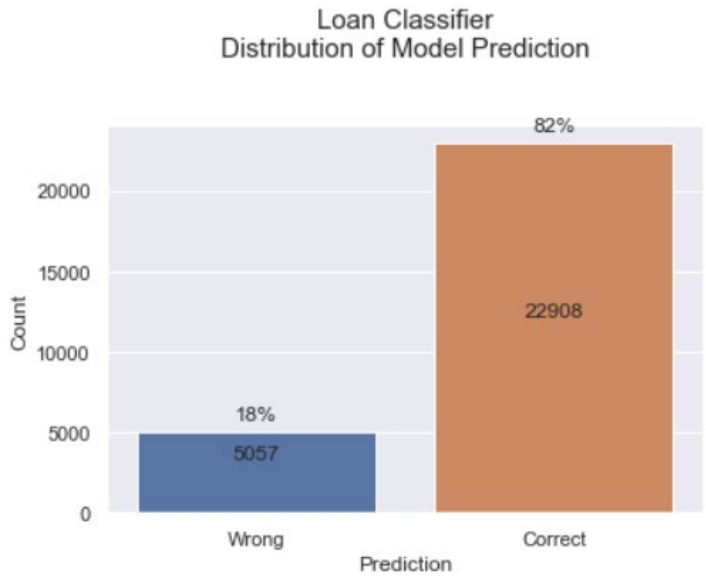
The feature important for model 1 is illustrated below:



Below is a sample prediction made by model 1:



The plot below shows the distribution of model prediction:



Random Forest Regressor: Estimate Length of Loan

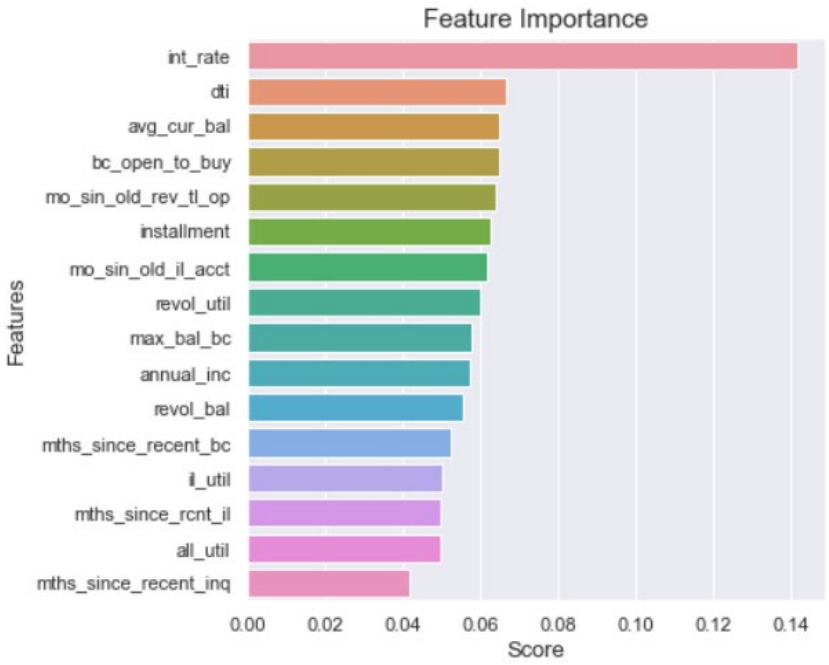
a. Model Evaluation & Selection

Base on the errors and R-squared value on the table below, base model seems to be the best model. However, since base model use all features in making the prediction, it is more likely to overfit. Model 2 (grid search) is the worst model because it has highest error scores and lowest R-squared. Therefore, model 1 (feature reduction) is selected in making the prediction because its' error scores and R-squared value are closed to that of the base model, and it also has the shortest training time.

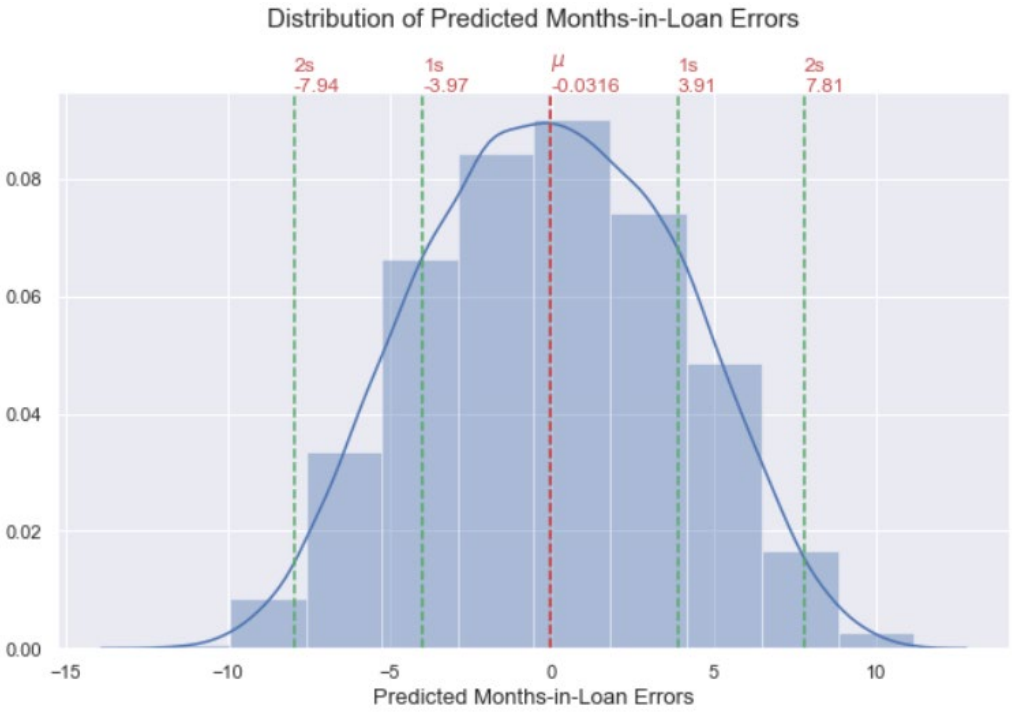
model_name	mse	rmse	mae	R-squared	training_time
base_model	15.385022	3.922374	3.237386	0.871849	240.061900
feature_reduction	15.510222	3.938302	3.246415	0.871484	87.549685
grid_search	16.500856	4.062125	3.373749	0.109093	338.585933

b. Model Prediction

The feature important for model 1 is illustrated as follow:



The plot below show the distribution of prediction errors. Most of the errors are closed to -0.0316. Since the distribution is normal, 68% of prediction errors are between -3.97 and 3.91, and 95% of prediction errors are between -7.94 and 7.81.



The table below show the predictions made by both the loan classifier and months-in-loan estimator:

loan_amnt	installment	actual_label	predict_label	prob_fully_paid	prob_default	actual_mil	predict_mil	predict_error	predict_mil_68cf	predict_mil_95cf
7000	210.78	fully-paid	fully-paid	0.88	0.12	7	9.9	2.9	(5.93, 13.81)	(1.96, 17.71)
1600	48.74	fully-paid	fully-paid	0.98	0.02	15.9	9.3	-6.6	(5.33, 13.21)	(1.36, 17.11)
20000	693.51	fully-paid	fully-paid	0.69	0.31	9	4.9	-4.1	(0.93, 8.81)	(0, 12.71)
7000	240.85	fully-paid	fully-paid	0.77	0.23	3	5.3	2.3	(1.33, 9.21)	(0, 13.11)
20000	684.33	fully-paid	fully-paid	0.84	0.16	14.9	7	-7.9	(3.03, 10.91)	(0, 14.81)
10000	275.34	default	default	0.47	0.53	9.1	7.6	-1.5	(3.63, 11.51)	(0, 15.41)
8000	251.4	default	fully-paid	0.97	0.03	1	4.2	3.2	(0.23, 8.11)	(0, 12.01)
10000	381.13	default	default	0.48	0.52	7	6.8	-0.2	(2.83, 10.71)	(0, 14.61)
6000	224.18	fully-paid	fully-paid	0.86	0.14	10	7	-3	(3.03, 10.91)	(0, 14.81)
15000	481.06	fully-paid	fully-paid	1	0	11	7.6	-3.4	(3.63, 11.51)	(0, 15.41)
8950	321.19	default	fully-paid	0.63	0.37	2	7.7	5.7	(3.73, 11.61)	(0, 15.51)
4800	154.71	fully-paid	fully-paid	0.96	0.04	0	7.7	7.7	(3.73, 11.61)	(0, 15.51)
4800	166.52	fully-paid	fully-paid	0.75	0.25	11	9	-2	(5.03, 12.91)	(1.06, 16.81)
30000	606	fully-paid	fully-paid	0.8	0.2	5	7.1	2.1	(3.13, 11.01)	(0, 14.91)
35000	745.03	default	fully-paid	0.79	0.21	3	5.2	2.2	(1.23, 9.11)	(0, 13.01)
4000	152.46	default	fully-paid	0.83	0.17	3	8.1	5.1	(4.13, 12.01)	(0.16, 15.91)

where

- ***prob_fully_paid*** = probability of fully paid
- ***prob_default*** = probability of default
- ***actual_mil*** = actual months-in-loan
- ***predict_mil*** = predict months-in-loan
- ***predict_erro***r = prediction error = predict_mil – actual_mil
- ***predict_mil_68_cf*** = 68% confident interval for predict months-in-loan
- ***predict_mil_95_cf*** = 95% confident interval for predict months-in-loan

Recommendation

According to the aforementioned analysis, to minimize risk of losing money, the investors should invest in a fractional part of a loan, and select a loan that has a 36 months loan term, low interest rate, installment amount, and low average balance of all accounts.

Jupyter Notebook

The hyperlinks to the python code for the project are as follow:

- Data Wrangling
  https://github.com/nphan20181/Loan-Default-Prediction/blob/master/loan_data_wrangling.ipynb
- 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
- Data Preparation for Model Development
  https://github.com/nphan20181/Loan-Default-Prediction/blob/master/lc_prep_data_for_ml.ipynb
- Random Forest Classifier: Loan Classifier
  https://github.com/nphan20181/Loan-Default-Prediction/blob/master/lc_random_forest_classifier.ipynb
- Random Forest Regressor: Months-in-loan Estimator
  https://github.com/nphan20181/Loan-Default-Prediction/blob/master/lc_random_forest_regressor.ipynb

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