Springboard Data Science Capstone Project
Lending club default prediction

Wei Dong

5/3/21

1. Introduction

Lending Club is a peer-to-peer (P2P) lending platform which is defined as the practice of lending to individuals or businesses through an online platform that matches lenders and borrowers. Lending Club has more than 3 million customers and their easily accessible historical datasets enable us to investigate the drivers of default in P2P lending.

On the lending club investors provide funds for potential loan applicants and earn returns depending on the risk they take. Therefore, accurate prediction of default risk in lending has been a crucial to minimize the associated risks.

This project aims to predict the default risk with machine learning model, which could help investors make better investment decisions and interpret the results based on the models.

2. Dataset and Cleaning

The data were collected from loans evaluated by Lending Club in the period between 2007 and 2018 (www.lendingclub.com). The dataset was downloaded from Kaggle (https://www.kaggle.com/wordsforthewise/lending-club). Data dictionary for the Loan Stats is available to download from LendingClub website (https://help.lendingclub.com/hc/en-us/articles/216127307-Data-Dictionaries).

The dataset comprises 2.2 million rows and contains 151 features. Loans which have the status of fully paid or defaulted were used as target label for default prediction. We first removed the features that are not available before loan initiation. Based on investigation of the missing data, we dropped the features based on their missing portion or irrelevance to loan default. The remaining missing data was imputed before building a model.

3. Exploratory Data Analysis

From the dictionary for the features, we get an idea of what features are available at the point of loan application to avoid data leakage. After removing the irrelevant features, we have found that additional 35 features missed more than 30% of the data. To identify the pattern of all missing data, we have found two groups of features having the similar missing pattern (figure 1). We simply removed features in group2 because their average missing percentage is 95%. The remaining features not contributing default prediction were also dropped. Another data cleaning was to drop the records with unknown loan issue date, term, and loan status.

Loan status is our interests, and it is shown 79% of loans are fully paid and 21% of loans is charged off (figure 2). The borrowers with longer employment length have shown lowest default rate. Therefore, we might infer that unemployed borrowers are very likely to be charged off (Figure 3). Although most of loan types are individual loan, loan with joint application types have a higher default rate (Figure 4).

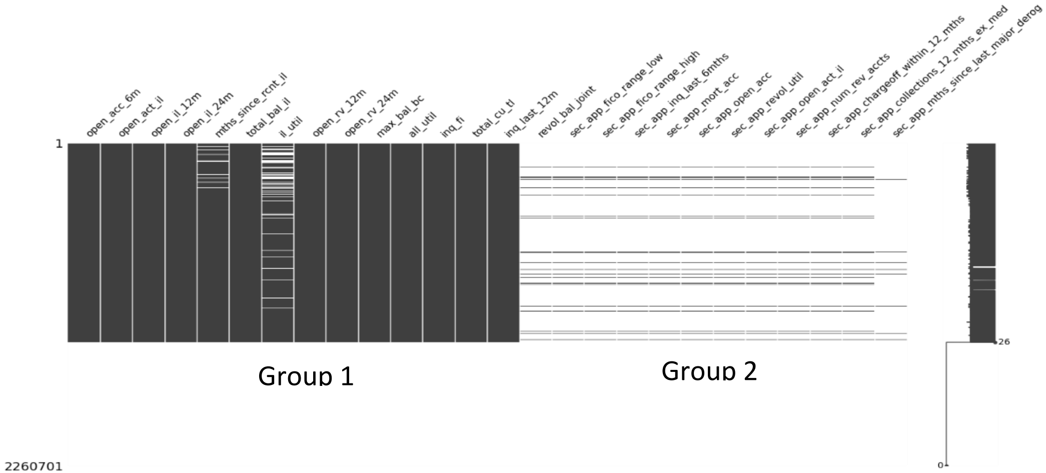

Figure 1. The pattern of missing values in features

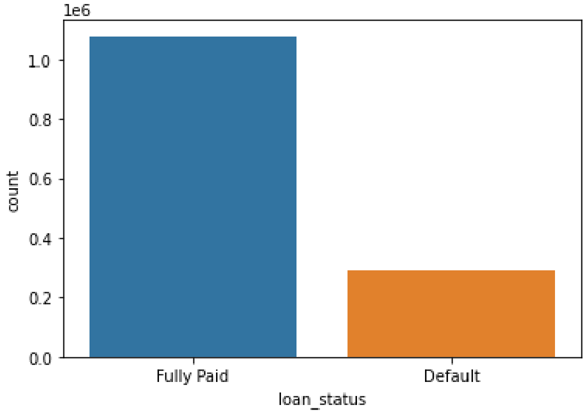

Figure 2. Loan status distribution

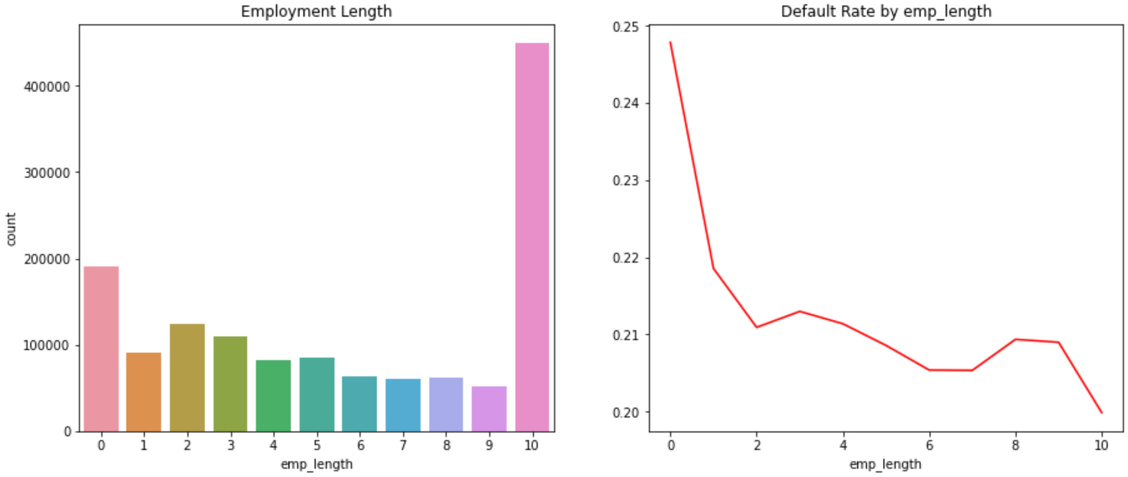

Figure 3. Employment distribution and their default rates

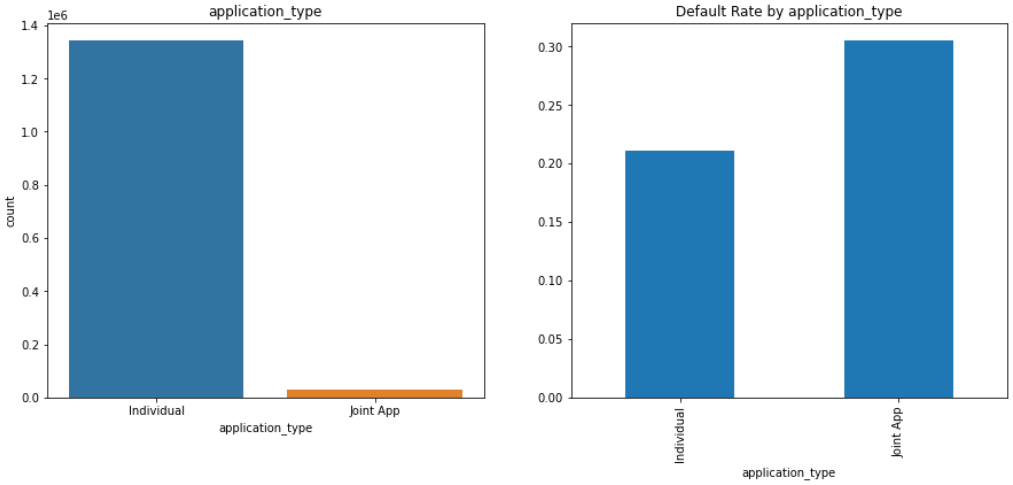

Figure 4. Loan application types and their default rates

We also wanted to verify that if low credit scores or low subgrade defined by lending club would cause higher default rate. Clearly, we can see that default rate is going up with the order of credit grades. And most of the borrowers have a grade B and C (Figure 5).

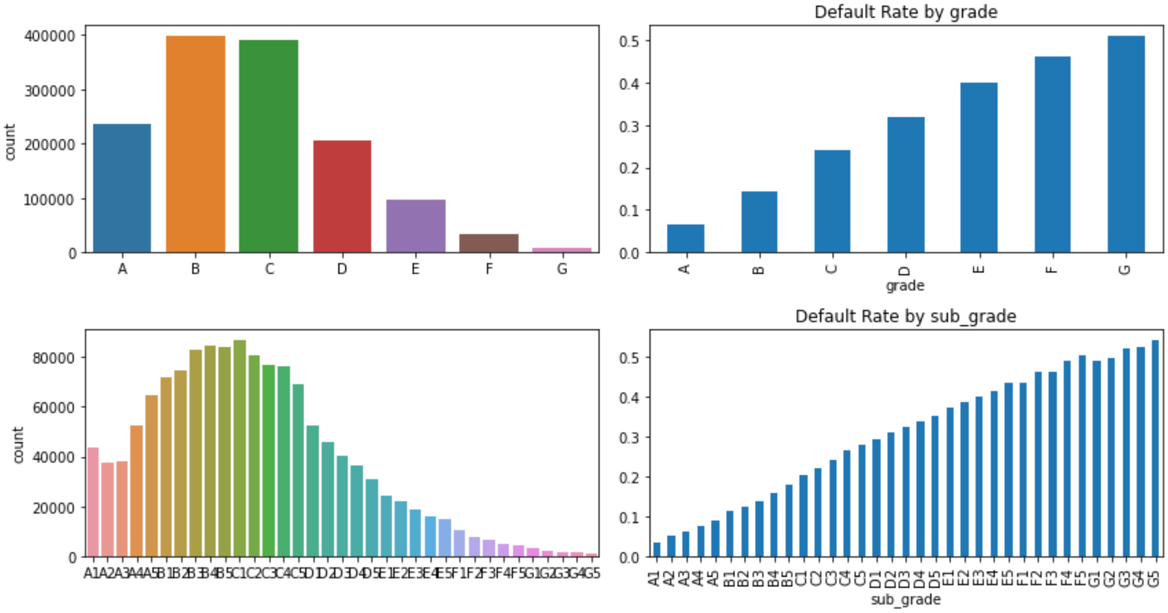

Figure 5. Loan application types and their default rates

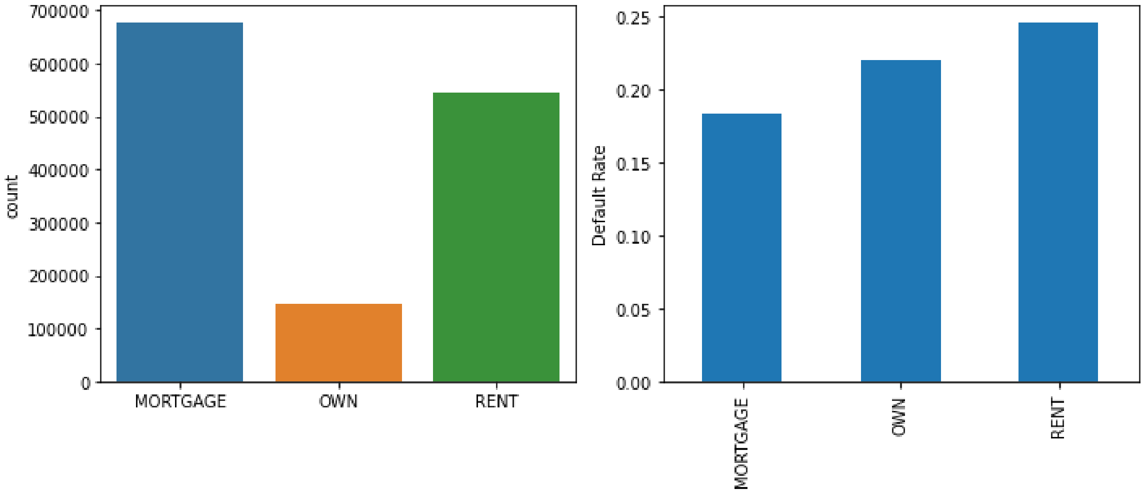

Figure 6. Homeownership and their default rates

Borrowers who live on a rental caused a higher default rate, and it is shown that borrowers with less stable living status are likely to default (Figure 6).

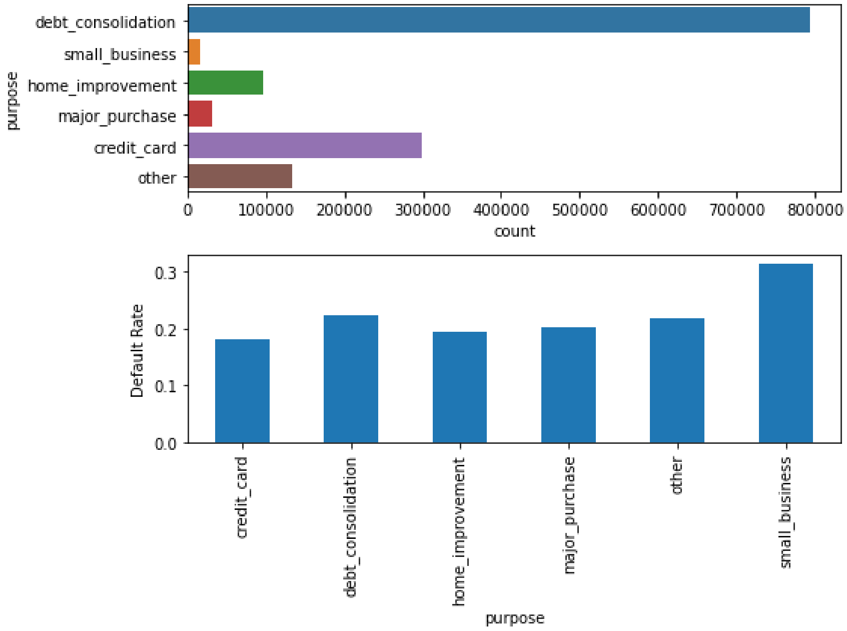

Figure 7. Loan purpose and their default rates

The loan purpose provides information about the purpose for which the loan was requested. Although most borrowers applied the loan for a debt consolidation, it is particular interest that small business was observed to have the highest fraction of the default rates (Figure 7).

Where are the loan applicants located in the US? We clearly see that most the California contains the largest fraction, but Mississippi state shows a relatively higher default rate (Figure 8).

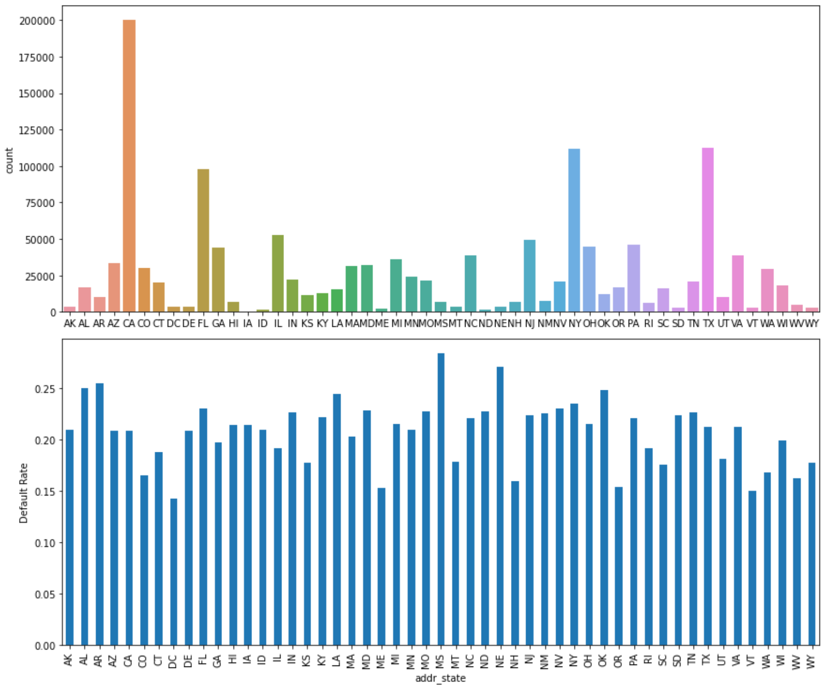

Figure 8. Loan applicant residence and default rates

Does the loan amount would cause higher default? Does only high amount of loans associate with higher default?

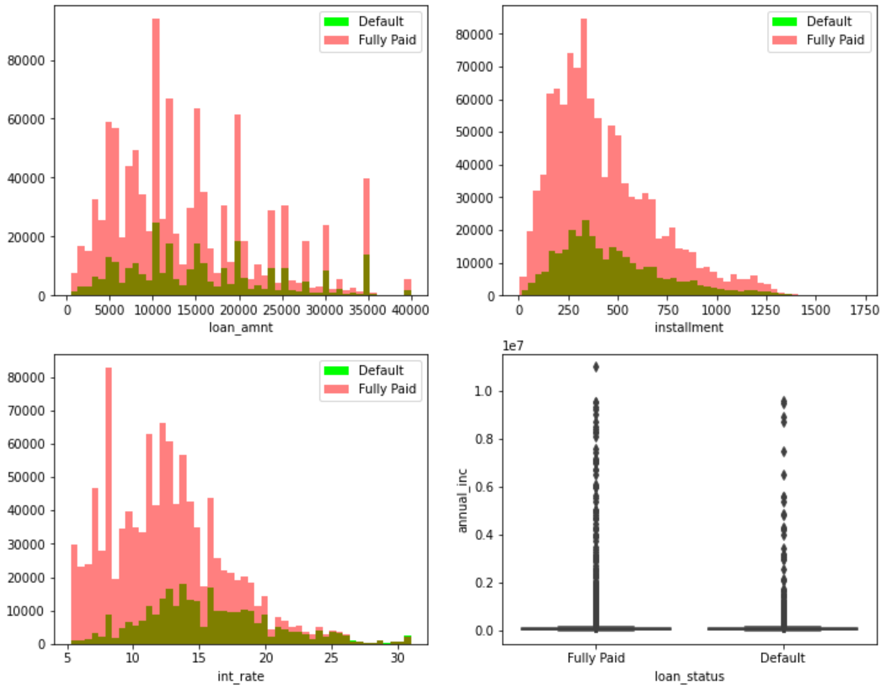

Figure 9. Loan status distribution with different loan amounts

Most of borrowers applied the loan around 10k $. Loan with different amounts demonstrated that a possible default. High interest rate has associated with default. While the borrower has relatively low credit scores, at some extent, they would be facing higher interested rate. All related information may cause loan default (Figure 9).

Another important factor may affect default is the Debt-to-Income ratio. An extreme high debt to income would not be realistic to get approved or get paid off. There are extreme data of dti ratios, in our case, we have removed the dti ratio that is higher than 800 (80%) (Figure 10).

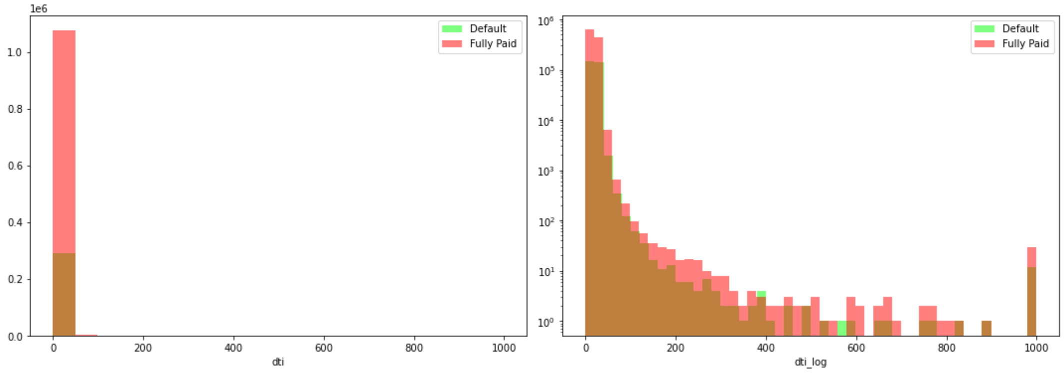

Figure 10. Loan status distribution with different loan amounts

Since the loan data was from 2007 to 2018, we should look at some basic information the loan through out the years. We have seen that lending club has issued much more loans than the other years, where 2015 has the peak. And the loan amounts started increasing quickly from 2013 to 2016. After showing a deep drop in 2016 and kept declining since 2016 (Figure 11).

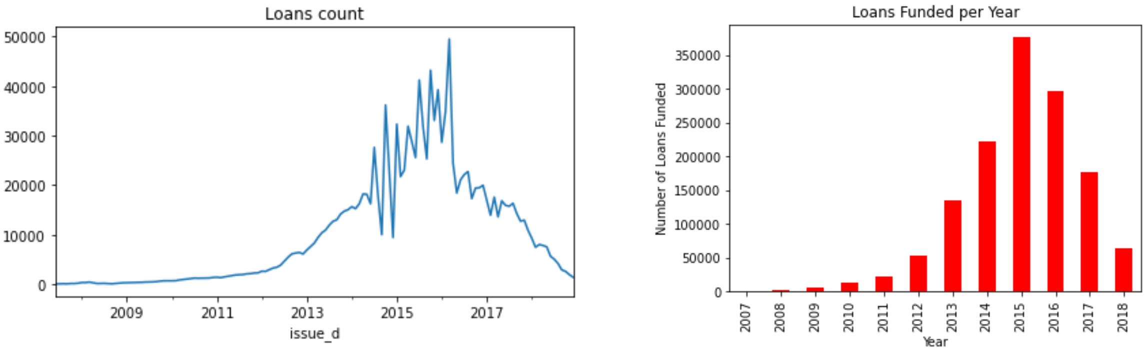

Figure 11. Loan status distribution with different loan amounts

The correlations among the features demonstrate some features having highly correlations that we want to take a close look at their relationships (Figure 12, Table 1). The top correlations are listed in table 1, and in this case, we have dropped one of highly correlated features. The major purpose was to avoid multicollinearity and an temptation to use the most important features in our model prediction.

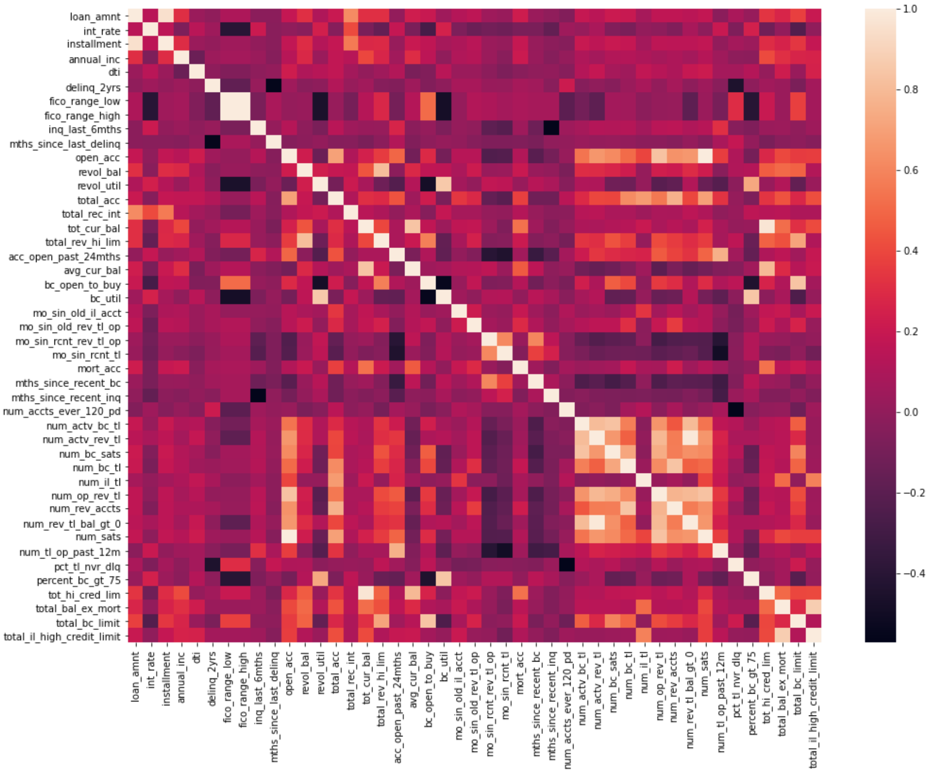

Figure 12. Heatmap with feature correlations

Table 1 Highly correlated features

Feature 1	**Feature 2**	**Correlation**
fico_range_low	fico_range_high	1.00
open_acc	num_sats	0.99
num_actv_rev_tl	num_rev_tl_bal_gt_0	0.98
tot_cur_bal	tot_hi_cred_lim	0.97
loan_amnt	installment	0.95

4. Modeling

We aim to build a model to predict the default of the loans. Therefore, there are two labels in the dataset (0 for fully paid off, 1 for default), we should use supervised machine learning algorithms and binary classification. The models are trained using 75% of the data and 25% of data for evaluating the model performances. There are 21% of the loans shown default status, hence we have a highly imbalanced data. There are three strategies to address the issue, resampling to get balanced data, decreasing the majority classes, and increasing the minority class. Additionally, we should use different models and measure the performance of the model using different types of metrics.

4.1 Data Pre-processing

It is necessary to pre-process the data before feeding into machine learning algorithms. We have applied different steps shown below:

1) **Dummy variables**: There is no categorical variables missing. Therefore, we don't need to crated NaN dummy variables. We will create dummy variables for categorical variables.
2) **Imputation**: we have missing values in numerical and ordinal variables. There are three types of imputation methods been applied in this case. Numeric values were filled with median values. The variables with less than 50 unique values, comparing to 1.3 million of rows, we imputed them with most frequent values. And the other missing variable are month related. They were imputed with arbitrary values (-1).
3) **Subset the data**: the default rate increasing every year since 2009. After 2007, there is a rapid decline of default rate. it was argued to be since defaults are a stochastic cumulative process and that, with loans of 36–60-month term, most loans issued in that period did not have the time to default yet. Exclude the 2018 would help avoid bias (Figure 13).

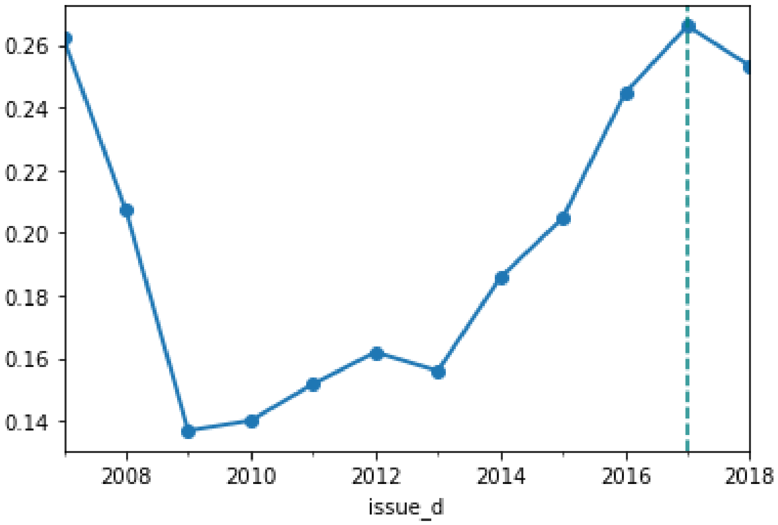

Figure 13 Default rates from 2007 to 2018

4) **Data splitting**: In this project we split the dataset into train and test datasets with a 75%-25% ratio.
5) **Weighting and undersampling**: it is time consuming to run large dataset, we decided to use undersampling strategy, upsampling and SMOTE should be also tried, which is not covered in this project. We decrease the non-default class to be identical with default label class, and then we use random sampling to use 0.1 fraction for model tuning.

4.2 Model

Two machine learning algorithms were applied to default prediction, logistic regression (LR) and Random Forest (RF). To reduce the local computing time, we used down sampling and random sampling strategies to build the machine learning model. We applied both basic model and model with optimized hyperparameters which shown improved metrics.

We want to have high true positive rate (or recall) with default group (class 1). And the model evaluation should also consider ROC AUC and PR AUC, in which AUC-ROC is the area under curve (AUC) of receiver operating characteristic (ROC) curve, and PR-ROC is the AUC of precision recall (curve). Technically, ROC curves are useful in an algorithm that optimizes PR AUC. It is reported that PR AUC is more informative for imbalanced data.

1) Logistic Regression (model)

We defined a LR model and performed gridsearch for best parameters. Based on best parameters, we found the model have good prediction of both classes (Figure 14, Table 1).

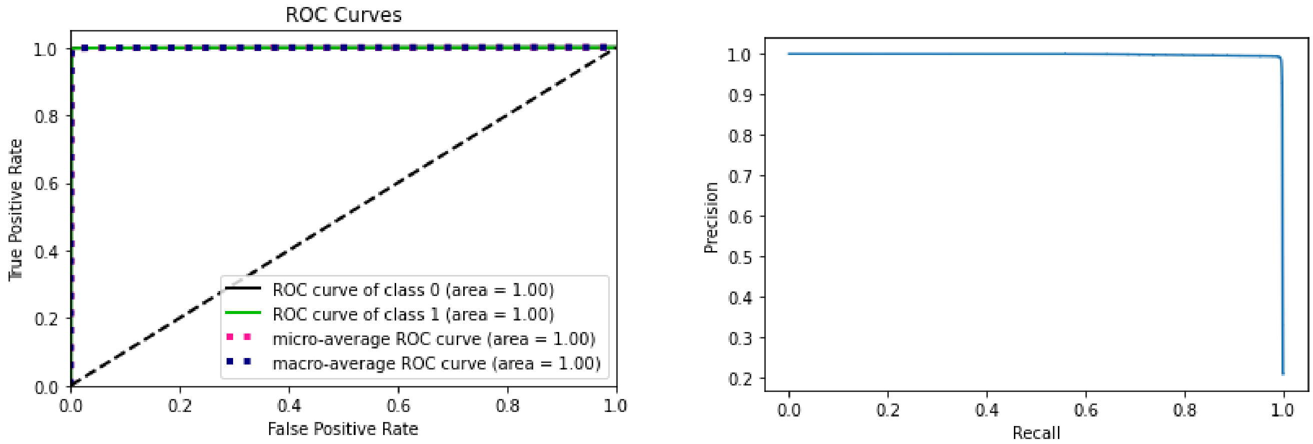

Figure 14 ROC and PR curves for logistic regression model

2) Same sampling strategy was applied in RF model. The AUC of base mode is 0.7057. We then selected a random group of parameters from pre-defined parameter grid and generated a tuned model, increase the AUC value to 0.7132. Although an exhausted random search of hyperparameter was performed and a propose best parameter was tested. There was neglectable increase. Therefore, the model with a random generated group of parameters was then used on unscaled data, and found that the AUC value has improved to 0.7208.

Table 2 Model metrics with select model

Model	Class	Values			ROC AUC	PR AUC	LogLoss
		Precision	Recall	F1-Score			
RF	0	0.89	0.60	0.72	0.3961	0.7208	0.6346
	1	0.32	0.72	0.45			
LR	0	1	1	1	0.9991	0.9973	0.0380
	1	0.99	0.99	0.99			

Overall, It is clearly shown that LR model has used less time than RF model, and have better model prediction for both class labels. Furthermore, additional model should be also tested, including Gradient boost model, and neutral network.