Who Will Pay the Loan?

A machine learning approach to identify good customers

Li Song

Part A: Exploratory Data Analysis

1. Dataset information

The raw dataset contains 161231 samples and 52 features. As checked, all the samples have unique id. Some features are numerical data, such as loan_amnt, funded_amnt. Some features are categorical data, such as grade, sub_grade. Some features like emp_length, int_rate need to be converted to numeric if they are considered in analysis.

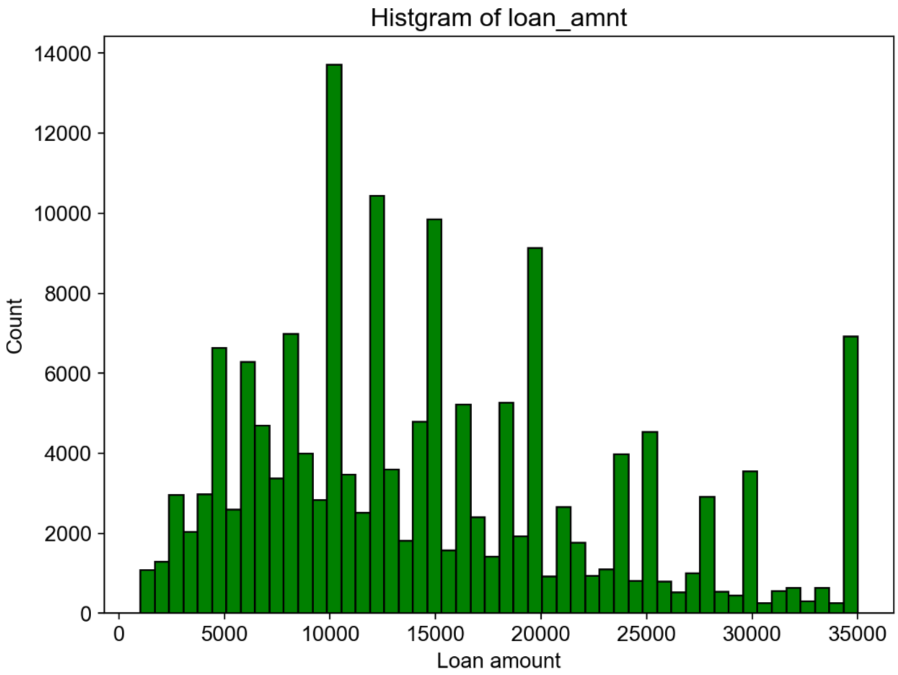
It is observed that there are missing values in the dataset. As computed, 9 out of 52 data features have missing values. The top 5 features that have the most missing values are shown below. Moreover, the features with missing values are removed from the dataset.

Feature	Fraction of missing values
desc	0.905744
mths_since_last_record	0.820574
mths_since_last_major_derog	0.719812
mths_since_last_delinq	0.497566
emp_title	0.056956

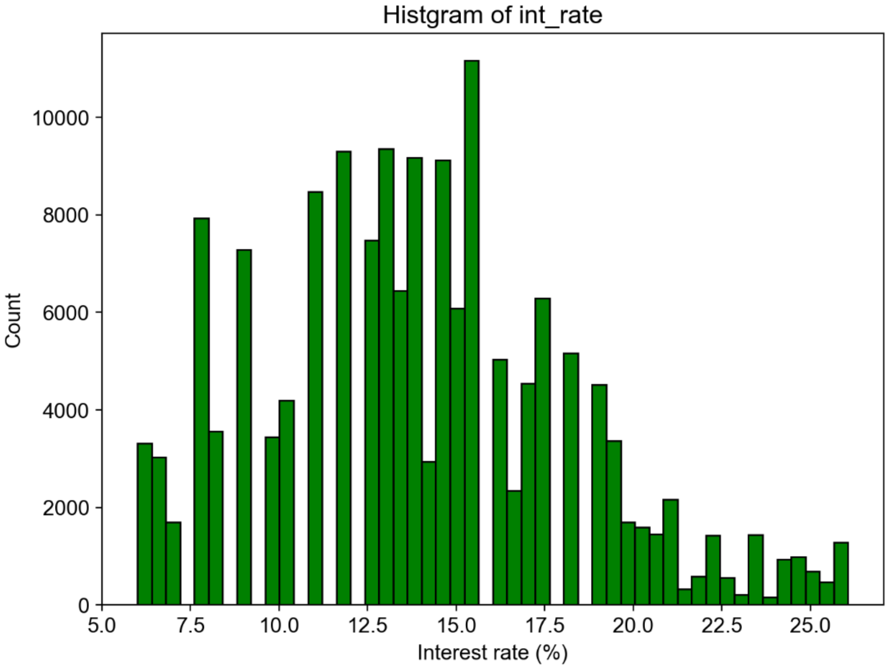
2. Descriptive statistics

In this section, three features loan_amnt, int_rate, and grade are studied. Loan_amnt, int_rate are numerical data, while grade is categorical data.

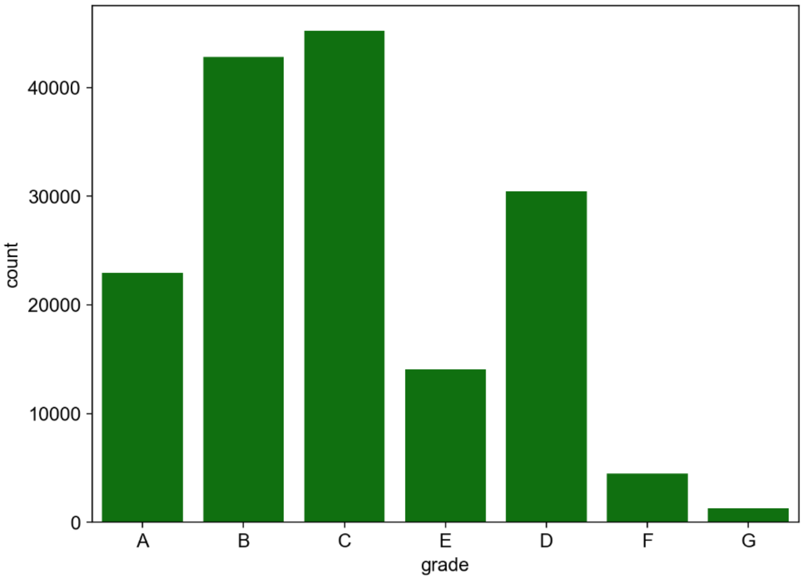
The minimum and maximum of loan_amnt are 1,000 and 35,000, respectively. The histogram of loan_amnt is shown below. It is seen that most of load amounts are lower than 20,000, and the highest frequency of loan is around 10,000. It indicates that most people prefer small amount of loan. Moreover, the data of loan amount is right-skewed.



The mean and standard deviation of int_rate are 14.06 and 4.33, respectively. The histogram of int_rate is shown below. It is seen that most of the interest rates are lower than 20%, and the highest frequency of interest rate is around 15%. Like loan amount, the interest rate is also right-skewed. It is obvious that higher interest rates come with fewer borrowers.



There are seven grades from A to G to be assigned to loans. Grade C has the most appearances, while grade G has the least appearances.



3. Business problem

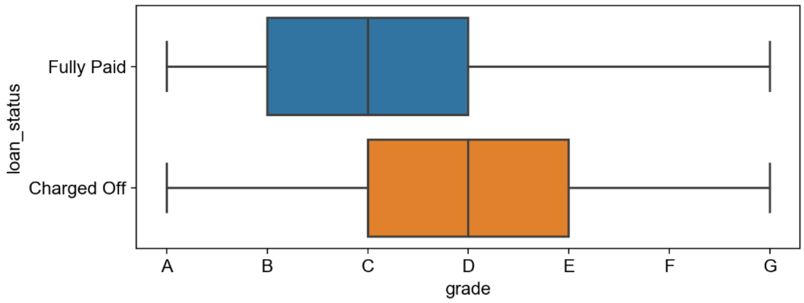
Now we look at the feature loan_status, which is a categorical data. The count numbers of each category are shown in the following table. My interest is paid to two categories: fully paid and charged off, and the rest are not considered in this section. The proposed business problem is that **Who Will Pay the Loan**? In other words, can we predict that whether a borrower pay or not pay the loan from the other features in the data? If the answer is yes, it will facilitate the decision-making process of identifying a good customer. Concretely, the company prefers to give money to customers who have high likelihood to repay. In such a way, the return on investment will be boosted, and the risk will be well-managed.

Category	Count	Description
Current	151208	
Fully Paid	6922	Loan has been fully paid.
Late (31-120 days)	1362	
In Grace Period	853	
Late (16-30 days)	409	
Charged Off	400	Further payment is no longer expected.
Default	76	
Issued	1	

4. Features related to the business problem

As defined in Section 3, the target feature is loan_status, which has two possible outcomes: fully paid and charged off. The category name cannot be used in analysis directly, so the dummy variables 0 and 1 are used to denote fully paid and charged off, respectively. In this section, two numerical features: loan_amnt and int_rate, and two categorical variables: term and grade are considered relevant to loan_status.

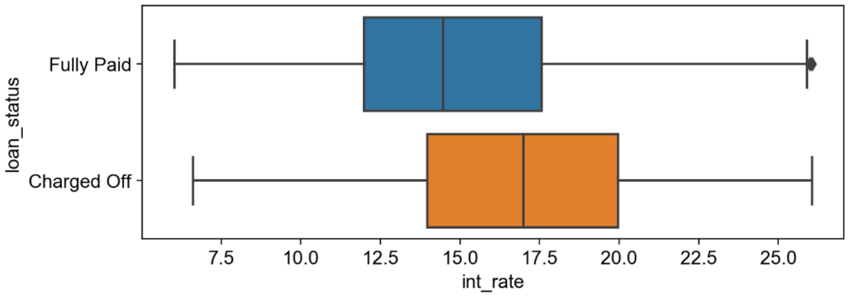
The feature grade takes 7 outcomes from A to G (i.e. from high to low). The box plot between loan_status and grade is shown below. On average, fully paid loans have higher grades than charged off loans. In other words, low grade loans tend to become charged off. For example, if the loan grade is lower than D, the likelihood of being charged off is higher than that of being fully paid.



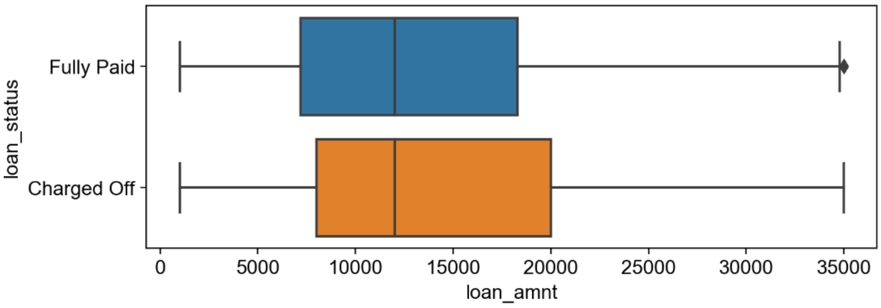
The feature term takes two outcomes: 36 months and 60 months. The statistics are shown below. For 36-month loans, 5.1% of them are charged off, which is lower than that of 60-month loans. It implies that long-term loans tend to be charged off.

Term	Total	Fully Paid	Charged Off
36 months	5238	4972 (94.9%)	266 (5.1%)
60 months	2084	1950 (93.6%)	134 (6.4%)

The box plot between int_rate and loan_status is shown below. On average, fully paid loans have lower interest rates than charged off loans. In other words, loans with low interest rate tend to be paid, which is consistent with our intuition. For example, when the interest rate is higher than 17%, the loan has a higher probability to be charged off.



The box plot between loan_amnt and loan_status is shown below. The distributions of fully paid loans and charged off loans are very close. Thus, given a loan amount, it is hard to tell which status the loan will go.



Part B: Feature Engineering and Modelling

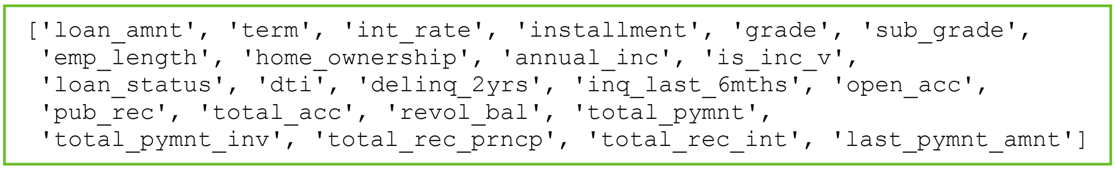
1. Machine learning model

a. Feature engineering

The number of features in the original data is 52. As pointed out in **Part A Section 1,** not all the features are useful in the analysis. Therefore, all the data features are reviewed based on the data dictionary and my knowledge, and some are removed with respect to the following criteria:

- Features with missing values, such as desc, mths_since_last_delinq, and emp_title.
- Irrelevant features based on intuition, such as id, member_id, and url.
- Redundant features, such as funded_amnt. For example, funded_amnt is the same as loan_amnt.
- Features with too many categories, such as purpose, title. For example, title has 2052 categories.
- Features with low variances, such as total_rec_late_fee, out_prncp.

As a result, the remaining 23 features are shown below..



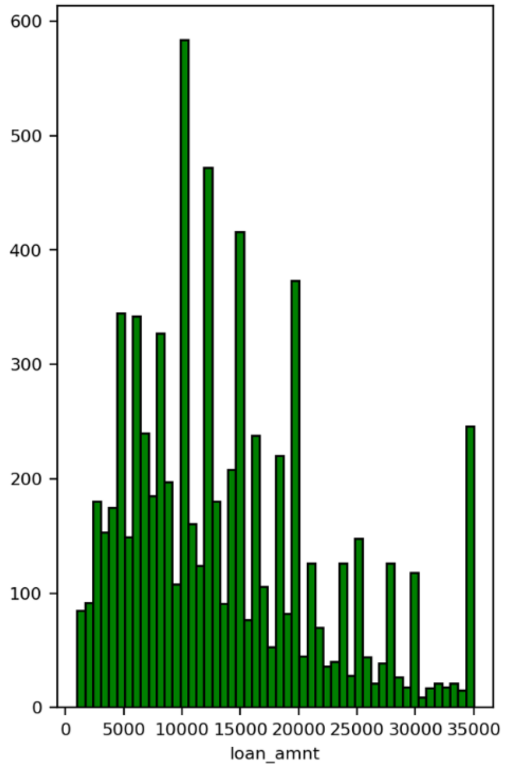
```
['loan_amnt', 'term', 'int_rate', 'installment', 'grade', 'sub_grade',
 'emp_length', 'home_ownership', 'annual_inc', 'is_inc_v',
 'loan_status', 'dti', 'delinq_2yrs', 'inq_last_6mths', 'open_acc',
 'pub_rec', 'total_acc', 'revol_bal', 'total_pymnt',
 'total_pymnt_inv', 'total_rec_prncp', 'total_rec_int', 'last_pymnt_amnt']
```

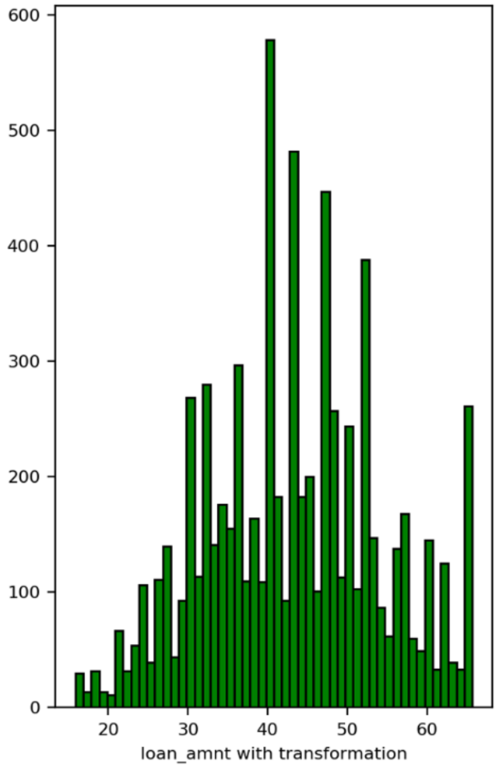
Two new features loan_amnt1 and annual_inc1 are created with the following transformations:

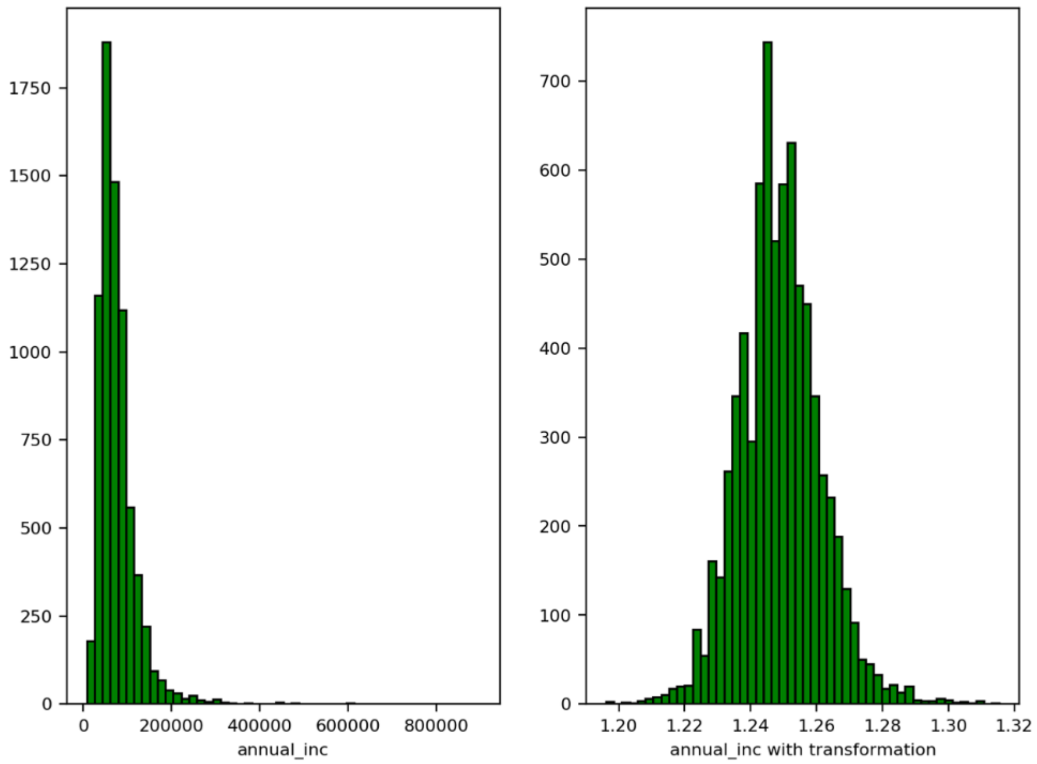
$$\text{loan\_amnt1} = \text{loan\_amnt}^{0.04}$$

$$\text{lnnual\_inc1} = \text{annual\_inc}^{0.02}$$

Generally speaking, machine learning models prefer data that follows normal distribution. Therefore, the rule behind transformation is to tune the data distribution towards normal distribution as close as possible. The histograms are shown below.





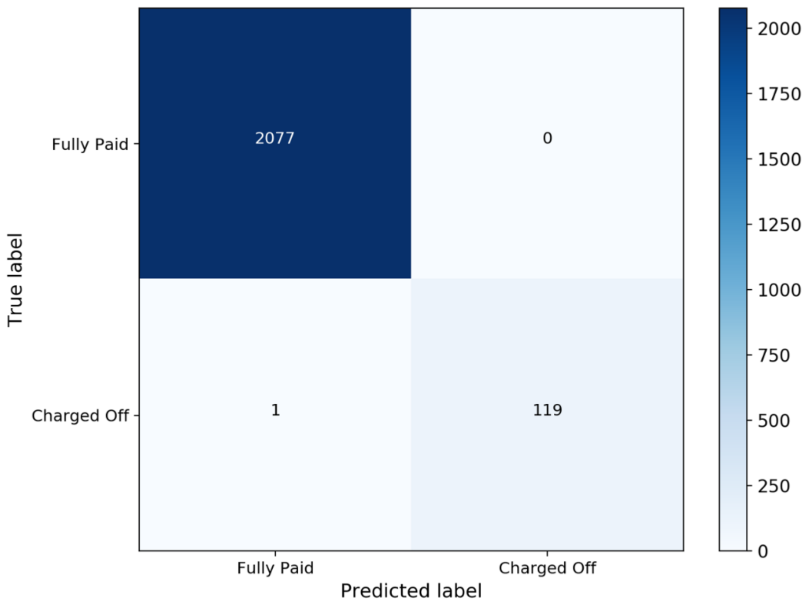


The new feature grade1 is created based on grade: the value is 0 if grade is higher than D, otherwise 1. The reason is that if the loan grade is higher than D, the likelihood of charged off is lower than that of fully paid. The feature home_ownership has four categories: mortgage, rent, own and any. We use 0 to denote mortgage and own, and 1 to denote rent and any.

b. Model development

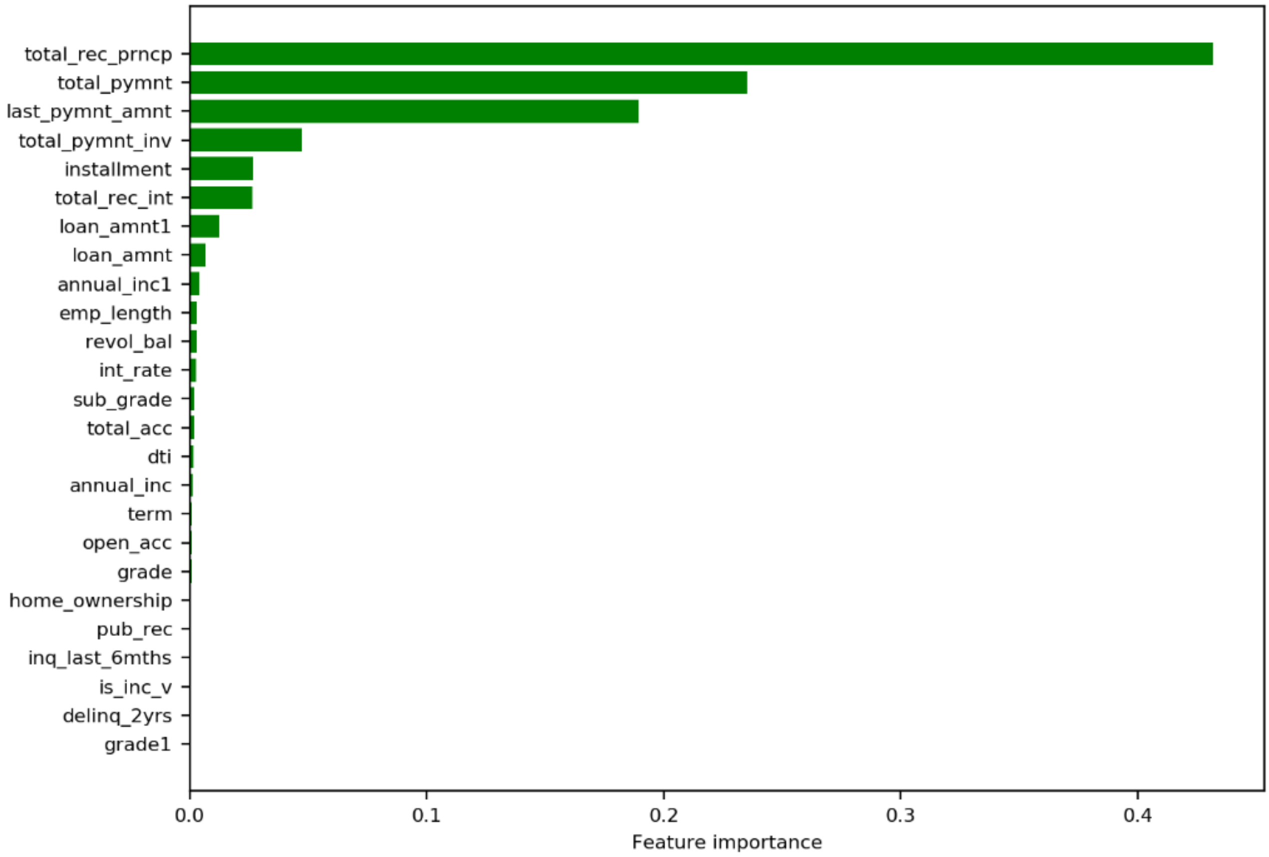
As proposed in **Part A Section 3**, the problem is to predict who will pay the loan. The dependent variable is loan_status, and the rest are the independent variables. Our task is to build a machine learning model to address this problem. As only two outcomes (i.e. fully paid and charged off) of loan_status are used, the proposed problem is essentially a binary classification problem. As we know, there are many off-the-shelf classification algorithms, such as logistic regression, decision tree, and support vector machine. In this section, **random forest** is our choice.

The data has 7322 samples in total, with 6922 fully-paid samples and 400 charged-off samples. It is seen that the data is highly unbalanced, which increase the difficulty of this problem. As a common practice, 70% of data are used as training set, which are randomly chosen. And the rest 30% are used as testing set. The random forest model comes with default parameters. The confusion matrix of prediction result is shown below.



It is seen that the model performs perfectly on the proposed problem. The prediction accuracy is 99.95% and the F1 score is 99.58%. Only 1 charged off sample is wrongly predicted as fully paid. As the model performance is already extremely good, it is not necessary to conduct feature selection and grid search. Moreover, some specific treatments on imbalanced data can also be skipped.

One benefit from random forest model is that it can quantify the importance of features. As shown below, total_rec_prncp, total_pymnt, last_pymnt_amnt are the top three in terms of feature importance. Moreover, the newly generated features loan_amnt1 and annual_inc1 are more important than their original forms. It indicates that the transformation towards normal distribution is effective.



2. Recommended feature set

As discussed in Section 1a, the number of data features is reduced from 52 to 23, in which loan_status is the target feature. Note that the removal is based on the exploratory data analysis and my understanding on the data dictionary. Some suggestions on the second-round analysis are given below:

- Consider the features with many categories, such as purpose, title.
- Merge categories to reduce the number based on the domain knowledge.
- Consider one-hot encoding for categorical data.
- Conduct transformation to other numerical data.

Some feature selection techniques such as sequential feature selection can also be involved in this case. Bear in mind that selection process should be based on cross validation.

3. Summary

In this report, I have studied the loan data, and proposed an interesting business problem: can we predict that who will pay the loan? The data have been wrangled step to step to meet the requirements of machine learning model. The relations between loan_status and loan_amnt, int_rate, term, grade are reviewed.

As the proposed problem is a binary classification task, random forest model is selected. The prediction result is extremely good, even without additional fine-tuning. It demonstrates that the proposed machine learning model is able to predict that whether a borrower will repay the loan or not. Then we can believe that the proposed model can help the investor in decision-making and increase the return on investment.

As our insight is drawn from a small dataset (7322 samples), it is necessary to collect more data to verify. With new big size dataset, it is expected to take much more time to fine-tune the machine learning model. More auxiliary techniques such as feature selection, learning algorithms may be involved. However, the general working flow will not change.