Report of my methodology and findings

Ziyu Zong School of Computer Science and Cybersecurity
Communication University of China Email: ziyuzong@cuc.edu.cn

Background and Purpose

- There is a lending company that provides small loans to individual borrowers for 3 or 5 years.
- To conduct analysis on the data and uncover insights about the loan’s performance.

Experimental Environment

Operating System: Ubuntu 16.04

Program Language: Python 3.5.2

Pandas Version: 0.17.1

Process and Findings

Step 1. Data Processing

Based on the dataset, my data preprocessing is as follows.

Projects	Operation
first column	drop the column
funded_amnt	drop the column
mths_since_last_delinq	replace the N/A with number zero
emp_length	transform string to integer
term	transform string to integer
int_rate	transform string to float
N/A in rows	drop the row that has more than 5 columns with missing values
profit_or_loss	create a new column

Table 1: Data Preprocessing

According to the table 1, there are some details.

- Drop the first column and 'funded_amnt' (same as loan_amnt).
- According to the interpretation of 'mth_since_last_delinq' and the number of missing values, the borrower should don’t have last delinquency if the value is N/A, so I replace N/A with 0.
- For the same reason, the 'emp_length' is replaced by 0 if the value is N/A. In my method, the value is regarded as 1 if the values is ‘< 1 years’.
- Transform the string type into integer type. (term, int_rate)
- Create a new column named 'profit_or_loss' by 'total_rec_prncp' plus 'total_rec_int' minus 'loan_amnt'.

Step 2. Plot Diagram and Analyze

After data processing, I analyze features by SPSS and plot diagrams by Python. Lack of statistics and data mining theory, I tired a lot but can only find some normal phenomenon. There are some conclusion as follow.

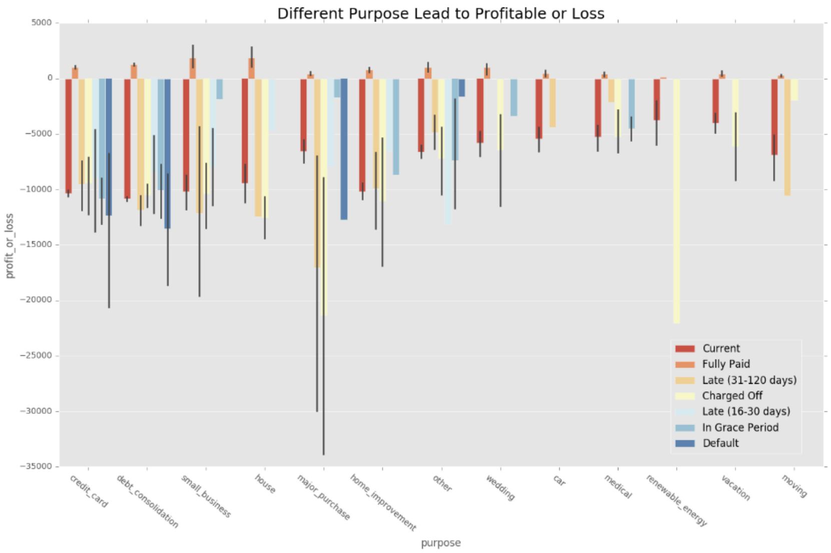

Figure 1: Different Purpose Lead to Profit or Loss

According to the bar chart which shows the average profit or loss, we could find the borrower who borrow and lend money for major purchase is most unreliable. Borrowers who make major purchasers are likely to cause big losses to the company. Among them, the most obvious ones are charged off and late (31-120 days).

It is most profitable that lend money for borrowers who have small business or borrow for house. Besides, they are relatively stable. So it is recommended that the company provide loans to people who have the intention to buy houses or who have small business and not to people who want to make major purchase.

To analyze the correlation about the features, I plot a heatmap.

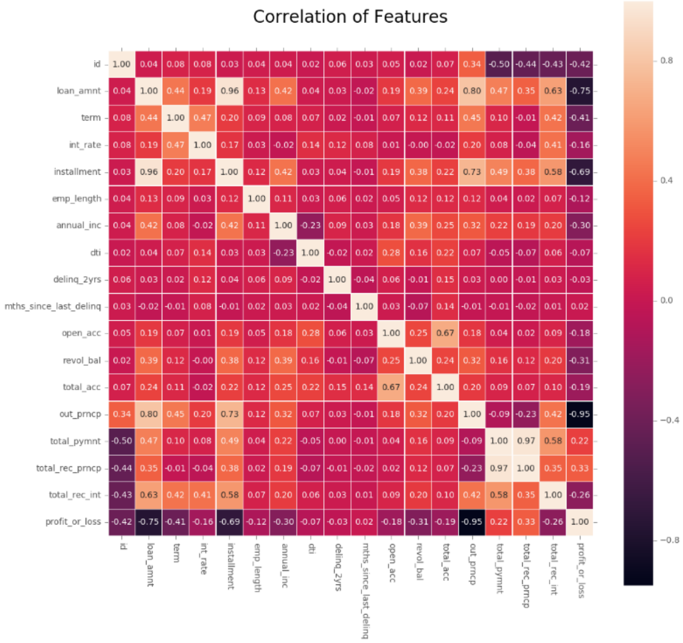

Figure 2: Heatmap about Features

In this figure, squares that are too dark or too light show the features of row and column are collinear. I pick up some features to plot a scatter figure and analyze the correlation between them.

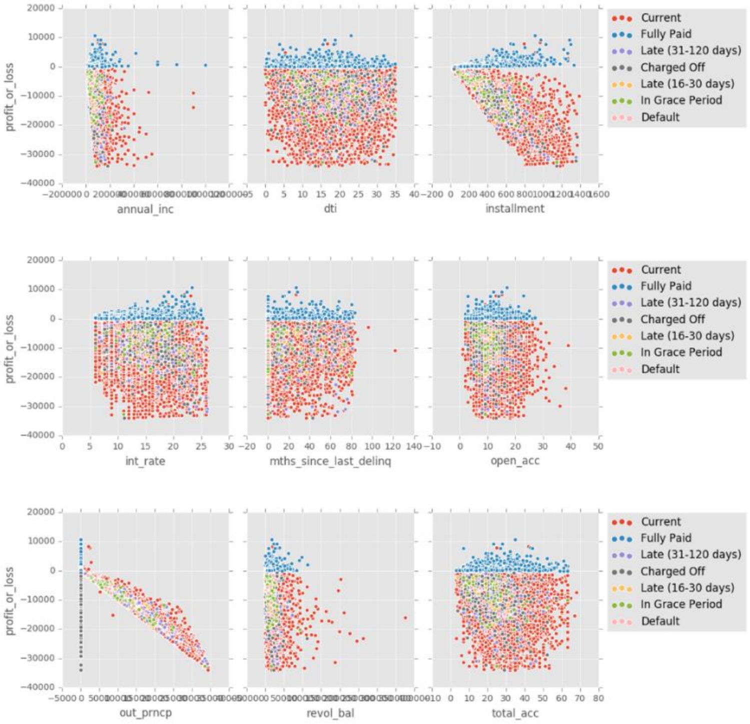

Figure 3: Scatter of Features and 'profit_os_loss'

In the scatter figure, the 'Current' dots are most quantity and the 'Fully Paid' dots are always profitable. Besides some normal phenomenon, I find a very interesting thing. The borrowers who have more than 20 credit accounts usually bring good loan, but people who have less credit accounts bring bad loan. Maybe some company have their own economic strategies through loan working capital.

Step 3. Using Model to Predict

According to the figure 3, I pick up some features to make further analysis. The features that I selected are as bellow.

'int_rate', 'installment', 'emp_length', 'annual_inc', 'dti', 'delinq_2yrs', 'open_acc', 'revol_bal', 'total_pymnt', 'profit_or_loss', 'loan_status'

After observing the data, I don’t think default should be used as bad label, so I filter out the data of the default label as the dataset to be predicted. And divide the data without default into train dataset and test dataset.

In the beginning, I used Logistic Regression and XGBoost to predict the default data set. Both of their accuracy is more than 90%, and their predicted result are all good loan. And I try to use some multiclassification models for predict, but the predicted result are all 'Current'.

After changing the model I got the same result, so I look up some reference and find this due to class-imbalance. There are 9073 good labels and 451 bad labels. The number of the good labels is 20 times than bad labels. So here are my two solutions.

I used the undersampling method to solve this problem. My first method is extracting some positive sample to make classes balance. Here is the number of samples I take for each classification.

Fully Paid	Current	Charged Off	Late (31-120 days)	In Grace Period	Late (16-30 days)
551	540	218	148	48	21

Table 2: The Number of Samples for each Classification

I use a neural network model named MLPClassifier (MLP) with two hidden layers, the first layer's size is 11 and the second is 2. Here is the Receiver Operating Characteristic curve (ROC), because the LogisticRegression (LR) and XGBRFClassifier (XGBRF) cannot product their predicted probability. So I can only get this figure.

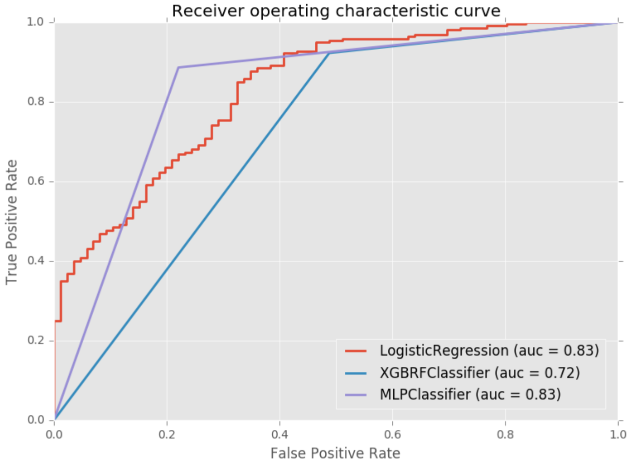

Figure 4: Receiver Operating Characteristic Curve

LR and MLP have large area than XGBRF. Both of their Area Under ROC Curve (AUC) are 0.83.

The second method I used is EasyEnsemble Algorithm, this algorithm use Ensemble Learning to divide the positive sample into many groups for different classifier. So it doesn't lose information on the whole. The accuracy rate of the EasyEnsembleClassifier (EEC) is 58.0%, and the AUC of EEC is 0.61. So it is not outstanding in accuracy.

	LogisticRegression	**XGBFClassifier**	**MLPClassifier**	**EasyEnsembleClassifier**
Accuracy	83.0%	80.7%	85.6%	58.0%
AUC	0.83	0.72	0.83	0.60

Table 3: The Number of Samples for each Classification

Here is the accuracy and AUC of these model in table 3. I think the LR or MLPC is better for the company to predict. The default model's prediction results thought different model is in figure 5.

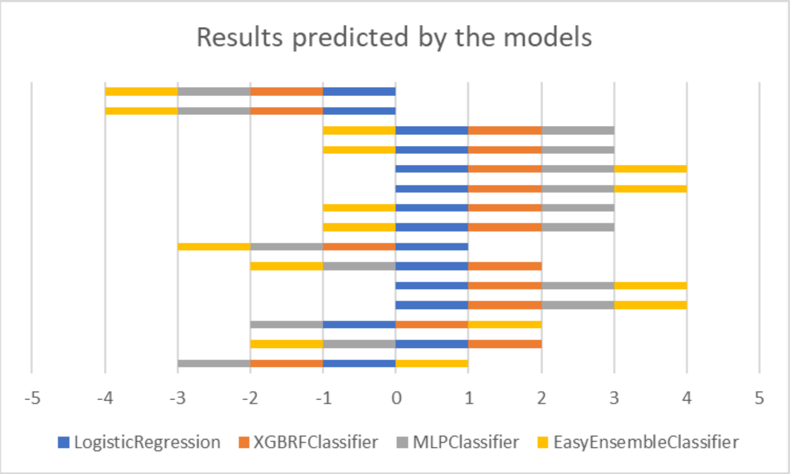

Figure 5: Results predicted by the models

In figure 5, the negative value means bad loan, the positive value means good loan. I put four model's prediction into one plot to predict the most probable label for the default loan status.

Thanks for reading. If you have any questions, please contact me. I will keep learning and try my best in our research.