Report of my methodology and findings in small loans

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
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
addr_state	map to four geographical division
profit_or_loss	create a new column
own_house	whether borrower has own house
mortgage_house	whether borrower mortgage house

Table 1: Data Preprocessing

According to the table 1, there are some details.

- According to the interpretation of 'mth_since_last_delinq' and the number of missing values, the bor rower 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 i s regarded as 1 if the value is '< 1 years'.
- Transform the string type into integer type. (term, int_rate)
- Map borrowers' address state to four geographical division (South, Midwest, West, Northeast).
- Create a column named 'profit_or_loss' by 'total_rec_prncp' plus 'total_rec_int' minus 'loan_amnt'.
- Create columns 'own_house' and 'mortgage_house', mean whether borrowers have their own house o r mortgage their houses.

Step 2. Analyze and Plot Diagram

At first I used the stepwise regression to fit a model, but its effect is not obvious. After data normal ization I use features 'int_rate','installment','emp_length','annual_inc','dti','delinq_2yrs','open_acc','revol _bal','total_pymnt','own_house','mortgage_house' to do Principal Component Analysis. The result show

s there are five principle components that is consistent with Credit 5C Evaluation. So I create 5 features and their details are as follow.

- Condition: 0.592 * revolving balance - 0.901 * amount of loan - 0.895 * monthly payment. It refers to the economic environment that may affect their ability to pay.
- Capacity: 0.626 * annual income - 0.521* interest rate - 0.404 * debt-to-income ratio. It refers to the debt paying ability of customers.
- Capital: 0.583 * revolving balance - 0.756 * debt-to-income ratio - 0.626 * opened credit accounts. It refers to the financial status of customers, indicating the background of customers' possible debt repayment.
- Character: 0.599 * total payment on the loan - 0.504 * delinquencies in the last 2 years - 0.391* months since last delinquency. It refers to the possibility that customers try to fulfill their debt repayment obligations.
- Collateral: 0.455 * owning house + 0.332 * employment length - 0.262 * mortgage house. It refers to the assets that can be used as collateral when a customer refuses to pay or is unable to pay.

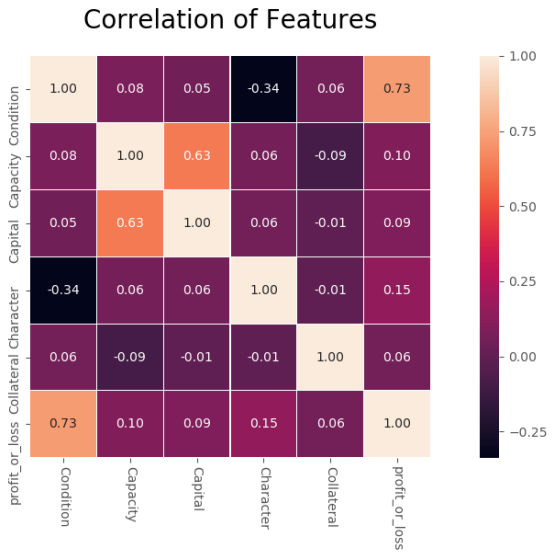

Figure 1: Heatmap about profitability and 5C evaluation

I plot a heatmap to analyze the correlation about profitability and 5C evaluation. In this figure, squares that are too dark or too light show the features of row and column are collinear. We can find the Character and Condition are more important than others in terms of profitability.

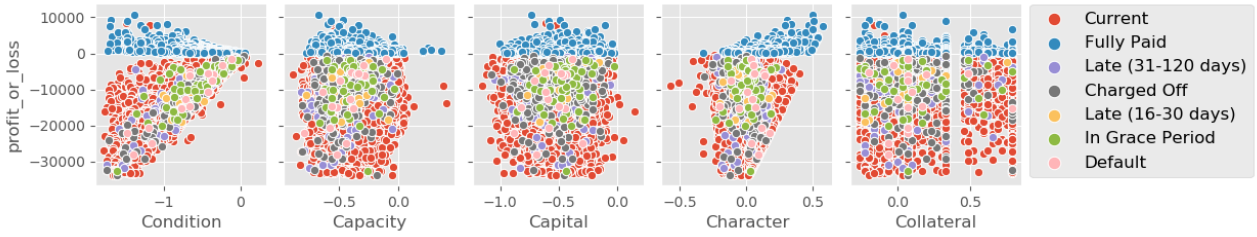

Figure 2: Scatter about profitability and 5C evaluation

In the scatter figure, we can see the distribution of the data and the 'Current' dots are most quantity and the 'Fully Paid' dots are always profit-able. We can find the same conclusion in the heatmap, the Character and Condition are more important than others in terms of profitability. So I plot a violinplot to explore whether there are differences in 5C evaluation model among borrowers in different districts.

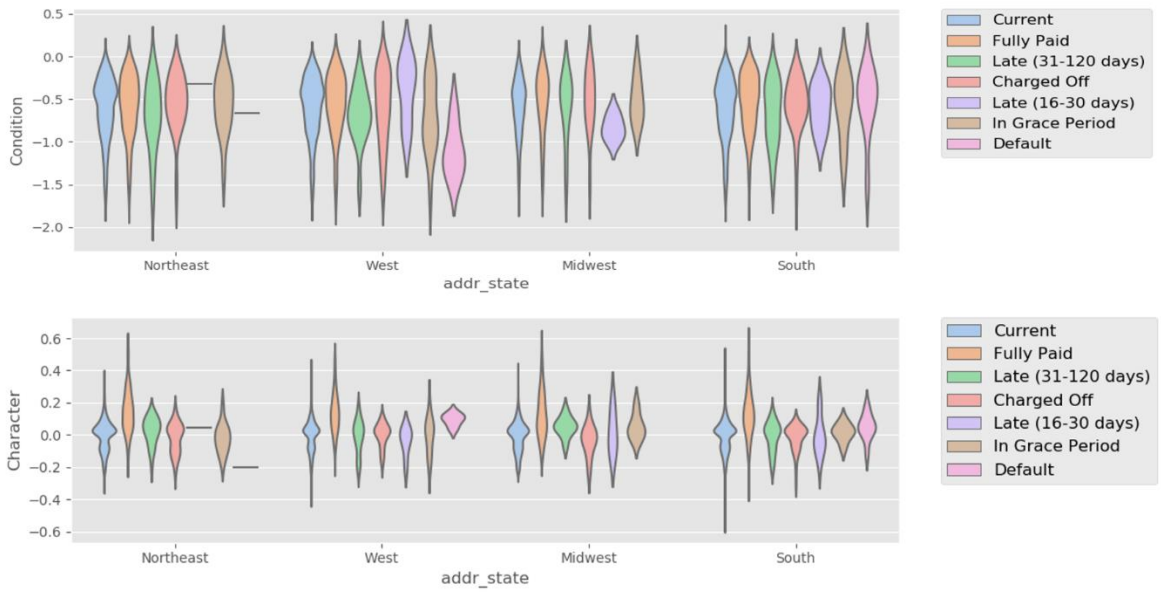

Figure 3: Violinplot about Address state, Condition and Character

It's obvious that Character is concentrated between -0.2 and 0.2, but Condition is distributed. It seems people in west have lower condition evaluation. The people in the four districts are almost the same. So the district is not the direct cause of the loan status.

Step 3. Using Models to Predict

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 training set(80% of the data) and testing set(20% of the data). I choose 5C evaluation as the input variables.

In the beginning, I used Logistic Regression and MLP 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 results 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 I used three undersampling methods to solve this problem.

My first method is using EasyEnsemble algorithm, this algorithm use Ensemble Learning to divide the positive sample into many groups for different classifier. Therefor it doesn't lose information on the whole. The accuracy rate of the EasyEnsemble is 58.0%. It is not outstanding, so I abandon this method.

My second method is extracting some positive sample randomly to make classes balance. Here is the number of samples I take for each classification. The number of positive samples is about twice that of negative samples.

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 three base classifiers. There are Logistic Regression, Decision Tree, MLP. Their optimizers are all quasi-Newton method. I chose 1 hidden layer with 5 neurons and 2 hidden layers with 5 neurons in first layer and 2 neurons in second layer. The activation of MLP is relu, the rectified linear unit functi

on (return $f(x) = \max(0, x)$). The initial learning rate I used is 0.001. Besides, the criterion of Decision Tree is Gini impurity, it have not bad accuracy and precision, so I used XGBoost to ensemble learning based on Decision Tree.

I use 10-fold cross validation to estimate these models and chose two evaluation criteria, accuracy and precision. I calculate the mean and std of the 10-fold cross validation. My result is as follow.

	Logistic Regression	Decision Tree	MLP (one layer)	MLP (two layers)	XGBoost
Accuracy (mean)	0.7792	0.7937	0.8625	0.8304	0.7897
Accuracy (std)	0.0866	0.0749	0.0708	0.0953	0.0772
Precision (mean)	0.8127	0.8131	0.8808	0.8605	0.8176
Precision (std)	0.0219	0.0191	0.0246	0.0424	0.0412

Table 3: 10-fold cross validation result of the model after random undersampling

Usually, we can't have both of good accuracy and good precision, but we can find the accuracy(mean) of MLP over 80% and the precision(mean) of MLP over 85%, besides both of them have less than 0.1 std according to the table 3. Therefore, it can be judged that the MLP fitting is better.

Because randomly undersampling has low reliability. I use the third method. I compress the data by K-Means clustering. The positive samples are grouped into 800 clusters, the negative samples are grouped into 400 clusters. The center points of each cluster are selected as the new data set. 80% of them are selected as training set, 20% of them are selected as testing set. I used the same 10-fold cross validation and the same models in my second method. But the result is unsatisfactory, all of the models' accuracy are under 60%. I think maybe is the result of twice data compression.

According to the table 3, we can see that the fitting effect of one layer MLP is better than other obvious ones, and both the accuracy and precision are very considerable. And MLP is better than statistical model Logistic Regression. Therefore, one hidden layer MLP is proposed to solve the problem of small loans.

The prediction of the 'Default' data set with one layer MLP model after data compression is as follow.

[-1 -1 -1 1 -1 -1 -1 -1 1 1 1 1 1 -1 -1 -1] (sorted by id). There are 6 good labels (1) and 10 bad labels (-1).

Summary

Dear boss, I suggest that we can use 5C reputation evaluation standard to evaluate customers. I set five standards by principal component analysis. (Details are above) We can evaluate each customer based on the five evaluation standards.

We can use regression model or neural network model to predict the loan status of some users. We can decide whether to grant loans to some users according to this credit evaluation model. Besides, the district is not the direct cause of the loan status. So we can grant loans regardless of the district.