Machine learning Engineer Nanodegree
Udacity Capstone Project

Loan Default Prediction for P2P Lending

Andri Sumitro
April - October 2018

Domain Background

Peer-to-peer lending (P2P) is a form of online micro- financing that allows individuals to lend or borrow virtually without financial intermediaries and collateral. P2P lending has become very popular and is growing rapidly in a lot of country due to the ease in receiving credit without having to deal with any financial intermediaries, as we know that, applying a loan from traditional bank is a long and legit process as it requires demographics characteristics, historical payment data, credit bureau data, application data etc, which cause a lot of individual and small business has a difficulty to get loan from huge financial institution like bank, especially in developing countries like Indonesia, which represent a big market opportunities and huge potential to fintech company like P2P lending company.

The rise of the P2P lending is due to the low cost of borrowing money as they are operated online which keep the operation cost low and most of their processes are automatized by the system. Thus they do not have to hire people to manually take care of the process. In addition, it facilitates quick loan decisions as the operations take place on the Internet that connects borrowers and lenders instantly. Another reason which make P2P lending become more popular is because it allows borrowers with short credit history an easy access to credit with a lower interest rate than traditional banks. The cost is reduced from the unbundling of unnecessary services that are coupled with traditional intermediaries. Small scale borrowers, such as individuals and entrepreneurs and borrowers that are placed at the long tail of credit, are mostly attracted to P2P lending due to non-requirement of collateral for the loans and lack of financial intermediaries.

The algorithm that has been widely used in the past for the bank or P2P lending companies assess their customers' creditworthiness is logistic regression. In my capstone project, I would like to see if the recent classification algorithm like random forest and gradient boosting would outperform classical logistic regression and if that is the case, how descent the recent classification algorithm would be as opposed to logistic regression (Anne Krane, 2014).

The motivation of choosing credit scoring as my capstone project is because I am currently working in a fintech company and we are about to build a credit scoring model for entrepreneur. Therefore, I think this would be a good start for me before helping my company to build one, I deeply believe that I would definitely learn a lot and gain more insight from this capstone project, especially with the guidance, support and suggestion from Udacity.

Problem Statement

The evaluation process of online lending is simpler than the process in huge financial institution like bank. However, it accesses far more data than a traditional bank in making a loan decision and some research has show that the credit scoring model using machine learning algorithm that has been used by P2P lending are much more accurate than conventional bank and has the power of predicting if the loan is going to default or not as well as it also shorten the length of loan approval decision. However, great return always comes with greater risk, in comparison to conventional bank, P2P lending has much more higher default and fraudulent risks due to insufficient credit checking, inadequate intermediation, lack of transparency and the inherent financial status of typical online borrowers. Therefore, credit scoring prediction and management are very extremely important to P2P lending company..

The aim of this project is to build a machine learning problem which help to predict if a loan is going to be default or not before they finish paying back by using the data from lending club in Kaggle. Not only does it help to decrease the likelihood of default, but it will also increase the profit margin of the company. Moreover, this will definitely benefit a lot of individuals and small scale business, especially for those who has a problem of getting loan from bank and in fact has the ability to pay back the loan.

Datasets and Inputs

The dataset that will be used in this project can be downloaded at https://www.lendingclub.com/info/download-data.action, which provided by Lending Club and the date of the dataset has been used is from 2012 to 2016. The dataset contains 1279326 rows and 145 columns. The 145 columns include information such as borrowers' credit history (such as fico score), personal information (such as annual income, years of employment, zipcode, etc.), loan information (description, type, interest rate, grade, etc.), current loan status (Current, Late, Fully Paid, etc.) and latest credit and payment information. All the other details regarding the columns could be found in the dictionary for the definitions of all data attributes.

Solution Statement

The machine learning algorithm that will be used to predict loan default is logistic regression, random forest and XGBoost. As logistic regression is well known for predicting loan default in the past, so I would like to see if logistic regression will be still outperform the other machine learning algorithm. Before feeding all the data into the algorithm, some data preprocessing (removing columns, filling missing value, encoding categorical variable etc) has to be done as there are a lot of useless data that will cause the algorithm to generate a false output. This would help to speed up the training process and extract hidden pattern out of the raw data in much more accurate way. After we make a comparison if our algorithm perform better than benchmark model, we will also use the metrics that are described in evaluation metrics section to evaluate which algorithm performs the best.

Benchmark model

By looking at the distribution of the data, it showed that there are a lot of borrowers will repay the loan than borrowers will default the loan. Therefore, the benchmark/naive model that I am going to use in this project is a model that predict every borrowers will repay all the time. The model that I'm going to build using logistics regression, random forest and XGBoost must at least beat the performance of this benchmark model.

Evaluation Metrics

1. **Confusion Matrix**

A classic of evaluating binary classification problems is by using confusion matrix. There are four vital values in confusion matrix are following:

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

2. **Precision and Recall**

   i. Precision

- is the ratio of correctly predicted positive observations to the total predicted positive observations.

ii. Recall

- is the ratio of correctly predicted positive observations to the all observations in actual class.

iii. Accuracy

- is simply a ratio of correctly predicted observation to the total observations.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3. F1-score

Sometime it is a bit confusing when there are too many metrics that could help you decide which one is better model. This is where F-score come to rescue as it combines precision and recall into one single metrics. Basically it is computed using the harmonic mean of precision and recall, which will help us to find a balance between precision and recall, especially for the case there is imbalance class.

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

Project Design

1. Data Pre-processing

- Handle missing values and columns by either
  - removing the row or fill it with mean/median value

2. Feature Engineering

- Encode categorical variables into numerical variables using one-hot encoded
  - use ordered number if they are nominal
  - use dummy variable if they are not
- Remove some extreme outliers which may mislead the results
- Normalize the variable

3. Data Visualisation

- Explore the relationship between variables
- Plot the correlation matrix

- Get the overview of correlation of all variables and have a big picture of what kind of variable has the power to predict

4. Model Building/Evaluation

- Train our machine learning algorithm by feeding training set data
- Test the model by feeding test set to see their performance
- Compare which model is the best after fine tuning the model and see if these model are better than benchmark model.

References:

1. http://kldavenport.com/lending-club-data-analysis-revisted-with-python/
2. http://www.icommercecentral.com/open-access/peer-to-peer-lending-default-predictionevidence-from-lending-club.pdf
3. https://www.yifeihu.me/static/project.pdf
4. http://blog.yhat.com/posts/machine-learning-for-predicting-bad-loans.html
5. https://edoc.ub.uni-muenchen.de/17143/1/Kraus_Anne.pdf