A Study on China’s Online P2P Lending Platforms

by

Anpu Li

An honors thesis submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Science

Business and Economics Honors Program

NYU Shanghai

May 2020

Professor Marti G. Subrahmanyam Professor Guodong Chen

Professor Christina Wang

Professor Jens Leth Hougaard

Faculty Advisers Thesis Adviser

Table of Contents

[Anpu Li 1](#_Toc39361023)

[Abstract 2](#_Toc39361024)

[Introduction 2](#_Toc39361025)

[Part I: Data Analysis 5](#_Toc39361026)

[Interest Rate 5](#_Toc39361027)

[Transactions 8](#_Toc39361028)

[Regression Analysis 10](#_Toc39361029)

[Part II: Machine Learning Analysis 12](#_Toc39361030)

[Problem Statement 12](#_Toc39361031)

[Notations 13](#_Toc39361032)

[Model Criteria 14](#_Toc39361033)

[Solution Overview 16](#_Toc39361034)

[Data-level methods 16](#_Toc39361035)

[Algorithm-level methods 17](#_Toc39361036)

[Ensembles 17](#_Toc39361037)

[Result 18](#_Toc39361038)

[Future Work 20](#_Toc39361039)

[References 21](#_Toc39361040)

Abstract

Compared with traditional institutions such as banking, the emergence of P2P network lending has brought about a variety of new and flexible forms of lending. In P2P lending, instead of a complicated review process that is rooted in traditional lending, investors and borrowers can directly complete asset matching and mutual benefit. However, due to the lack of relevant laws, effective management, P2P platforms have potential risks such as withdrawal failures, investigation involvements, and runaways, which may cause extreme losses to investors, and this problem is particularly serious in China. Although there is abundant public data of P2P platforms on the Internet, they are not well structured and readily available for people to retrieve and make analysis. Therefore, this paper aims to propose an example model that uses data from Hongling Capital (“红岭创投”) and analyzes the risk in both econometric and machine learning approaches and from the perspectives of both the platform and the investors. My model catches the determining factors of loan defaults and helps both the platforms and the investors in understanding the current risk.

Introduction

As the control is gradually released on the interest rate by People’s Bank of China, the reformation to liberalize interest rates slowly draws to an end. Interest rates in the financial market will be more subject to market demands and new monetary policies. During the process of the liberation, Internet Finance, through the weakening regulation, has gained a wild surviving plant and experienced an unprecedented growth. Internet Finance has a crucial impact on China’s economic growth by greatly increasing the efficiency in financial activities, such as the widely-used mobile/third-party pay. When it comes to lending, Internet Finance not only satisfies the investment needs of the mass people, but partly solves the financing problems of small-to-medium enterprises (SMEs). Among the manifold Internet Finance modules, online peer-to-peer (P2P) lending platforms own the popularity. Below is the graph that shows yearly numbers of P2P lending platforms. (“P2P Online Lending Platform 2019 Annual Report”)

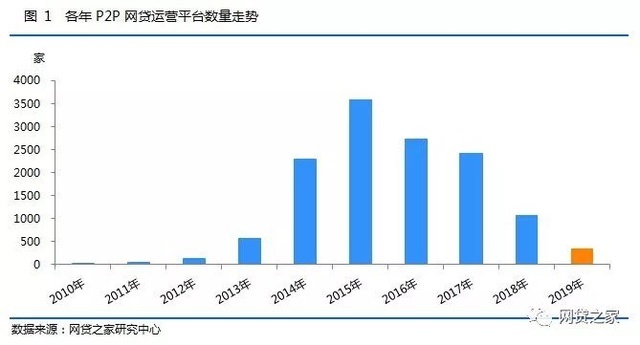


Figure 1. Number of Operating Platforms Trend

It is therefore practical to conduct more research on this typical Internet financial model. From the current point of view, due to the relatively short emergence of P2P online lending in China, the relevant laws and regulations are far from perfect. The market generally believes that the interest rate of the P2P online lending platform is excessively high and the risk is high also (He and Peng). However, high interest rates do not necessarily mean high risks, and low rates do not lead to low risks, and platforms with low interest rates also face potential risks such as bankruptcies owing to lack of check on the borrowers. The level of interest rate is not the whole picture. In recent years, P2P online lending rates have been declining due to the central bank's RRR cuts. The interest rate of the online lending industry is gradually returning to a reasonable range, which also reflects the significant impact of national financial policies and benchmark interest rates.

Due to the lack of effective supervision of the P2P online lending platform and the uneven development of the industry, most online lending platforms only use high yields to attract capital inputs but pay little focus on the risks behind. As a result, once one online lending platform collapses, it will bring widespread panic to market investors – and that is exactly what happened in the past years. Therefore, the improvement of the pricing mechanism of P2P online lending industry and the determination of interest rate level will be an important topic in the development of China's Internet finance. Moreover, the platform risk should be included as a crucial factor for investors before making investments, and for platforms before making expansions.

First, research on the changing trend of P2P online lending rates will help analyze the regulation of China's credit and its response to the benchmark interest rate of financial markets, because P2P lending rates, similar to private interest rates, are subject to the supply-demand influence and being liberalized. Second, studying the default behavior of borrowers will help the platforms to make right decision in times of rapid expansions. The rationality is important and indispensable in sustaining the platforms during the great crackdown of platforms in 2018, after outflow of hot money in the industry. More importantly, lots of people, especially the young people who have no ability to pay back, are attracted by the easiness of online borrowing and suffer from intolerable debts. Effective measures of preventing those people from getting the debt they do not necessarily and should not have will greatly help address bad loans and crackdowns.

In the paper, I 1) analyze the correlation and interdependent effects among SHIBOR, China bond yield, and P2P lending rates and conduct an event study on the influence of China’s stock market fall in 2015, and 2) propose a deep learning model based on data of more than 2 million users that captures the characteristics of defaults, and reaches a recall rate as a high as 93.6%.

Part I: Data Analysis

The main data used in the paper is credited to my advisor Professor Guodong Chen and his partner. The data is collected from my089.com (“红岭创投”), a leading P2P lending platform through the end of 2014 to the end of 2015. There are mainly two sectors of this data: one includes the bidder, bid amount, and time for every project; one includes the introduction of each project such as the interest rate, period, amount needed, and use. The latter will be the major focus on the paper. This paper also uses the Shibor data and the China bond yield collected from their official websites.

Interest Rate

Half products have an interest rate between 7% to 14%, with a mean of 10%. Interest rates underwent a significant decrease through the year, dropping from 12% to 7%, almost halved.

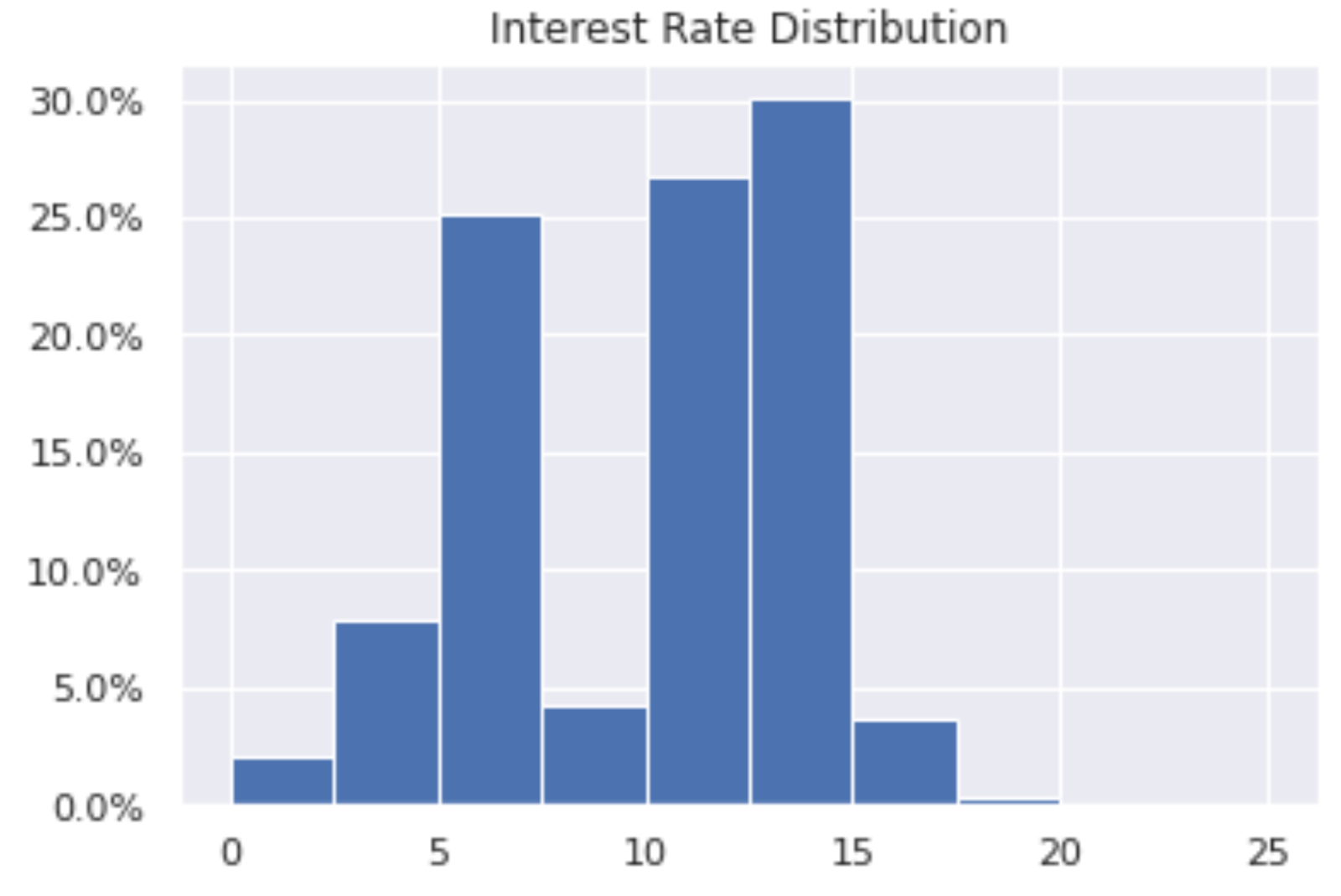


Figure . Interest Rate Distribution

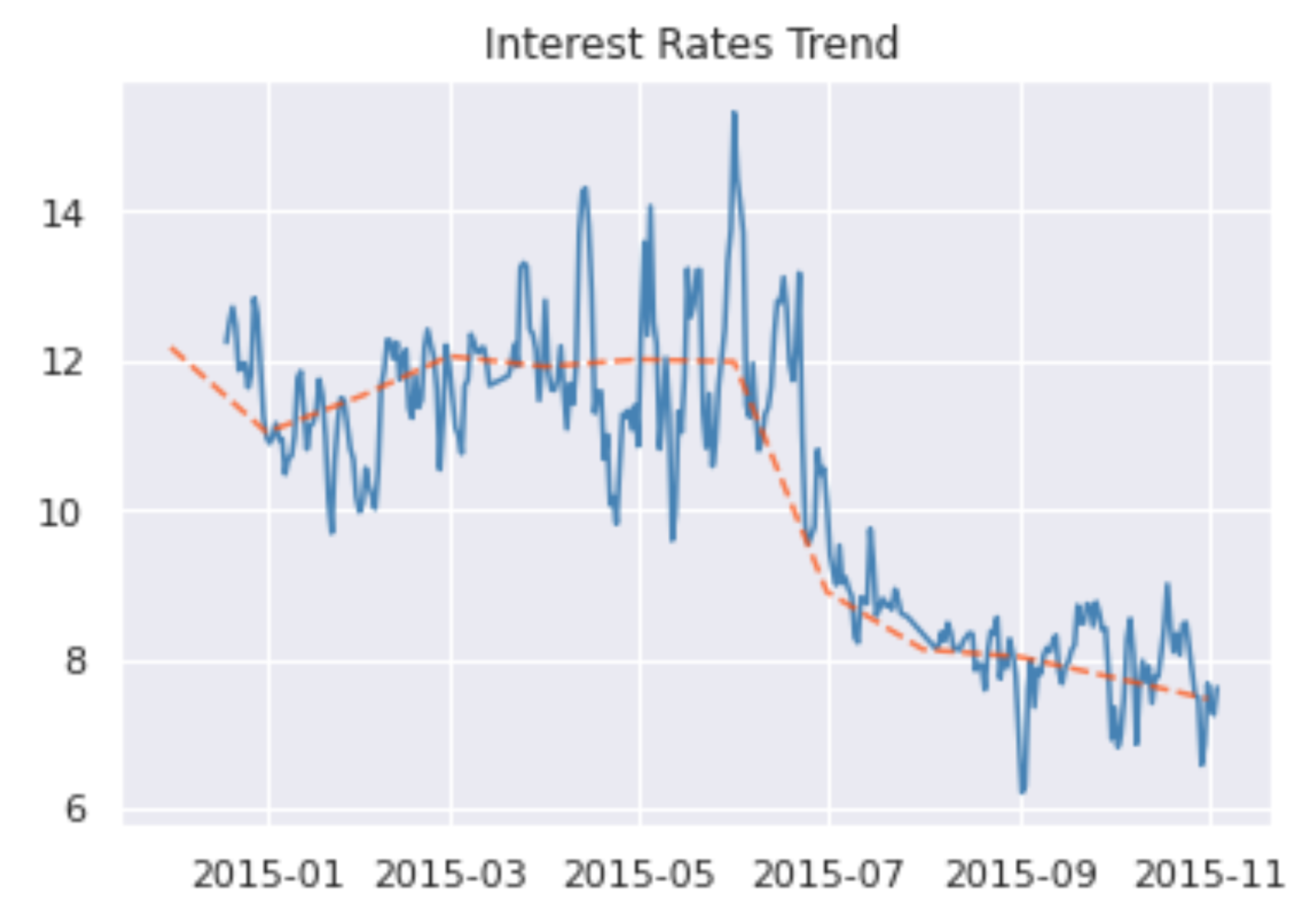


Figure . Interest Rate Over Time

The curves of the three rates are shown below. It is easy to spot that there is a positive correlation; however, during the stock market fall period, it seems to move opposite directions and has a long lag in responding to the market. A possible explanation is that the those seeing stock holders trapped in the market were pursuing ways to keep their cash in hand, therefore they are less supply on the online P2P platforms.

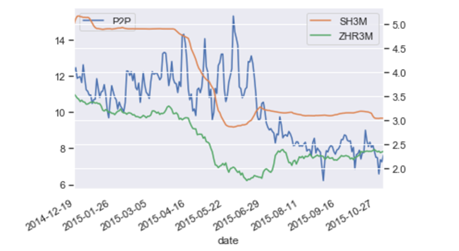
[](https://user-images.githubusercontent.com/26131764/70918819-aecc6e00-205a-11ea-8f97-18a16999c6dd.png)

Figure . P2P Lending Interest Rate Over Time

Transactions

The image below shows the number of projects over time. In the first half year, the curve is steady but in the second half, it slightly goes upward. One explanation could be because it is not profitable to invest in the market, people will go to other sources to make investment, and P2P lending is a popular and high-return product at that time. A close look at the distribution of loans, i.e., the loan state, points out the problem behind the seeming boom of P2P market. The majority of loans are recalled, a dangerous signal for the supply side.

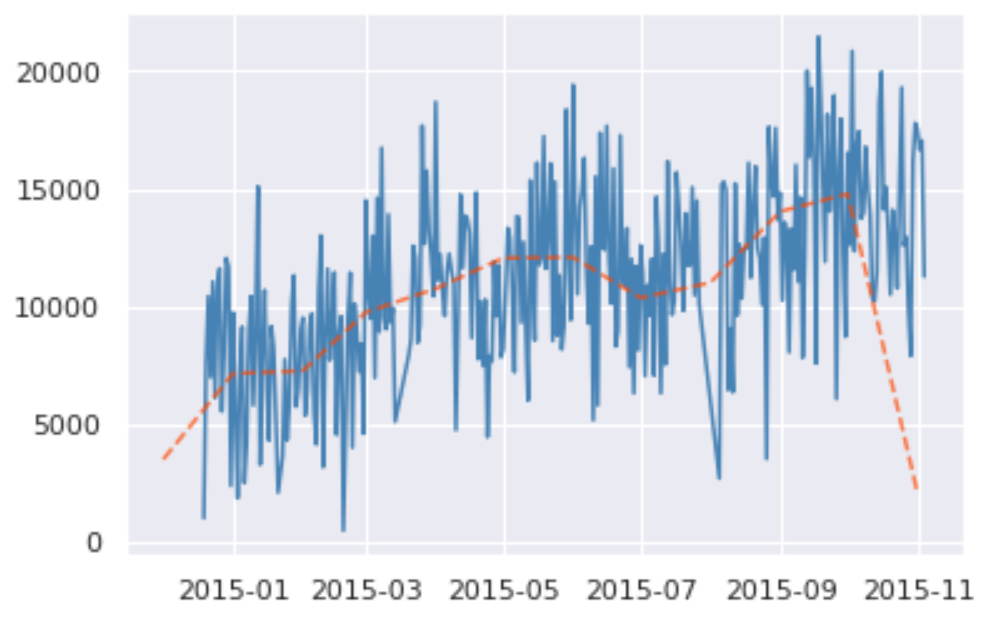


Figure . Number of Loans Over Time

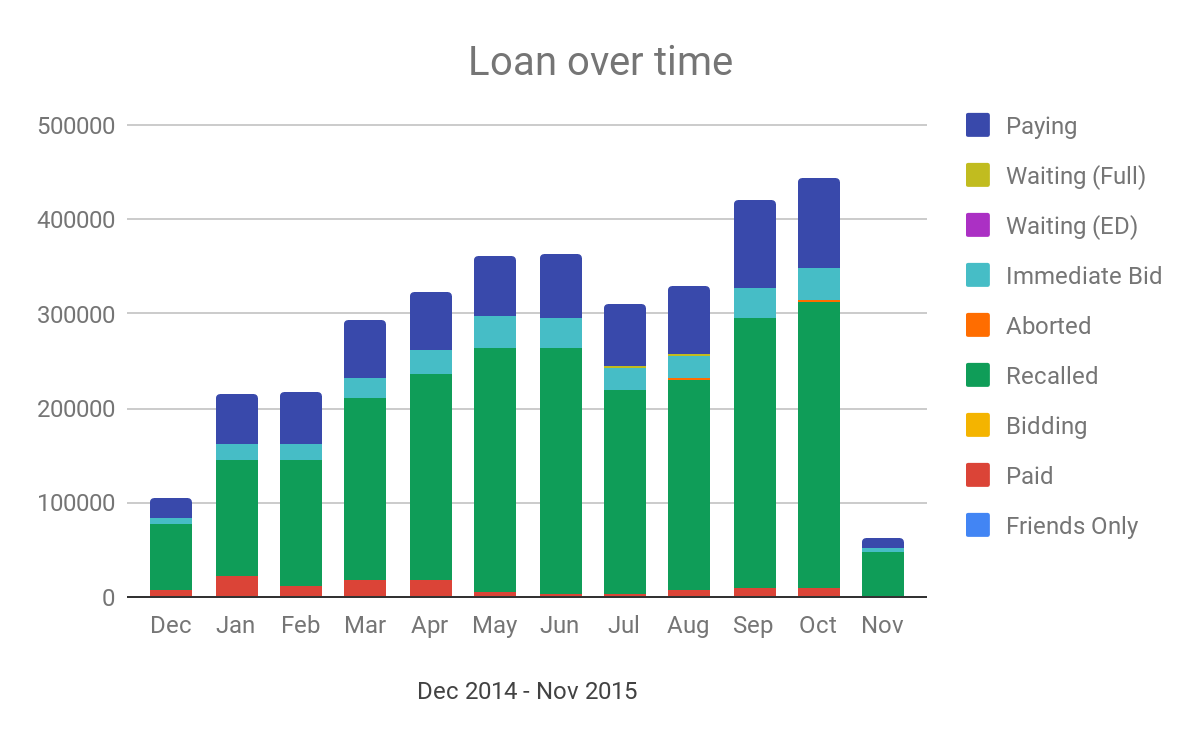


Figure . Distribution of Loans Over Time

The following is the overall default rate, it is clearly increasing with time, posing a red flag for lending money. But overall, the default rate was very low. There are mainly three explanations for this phenomenon. 1) The loans are mainly microcredit, so they are easy to repay. 2) The range of data is relatively short, so it may not contain all defaults in the limited data. 3) The duration of loans is short, so the platform takes actions to lower the default rates.

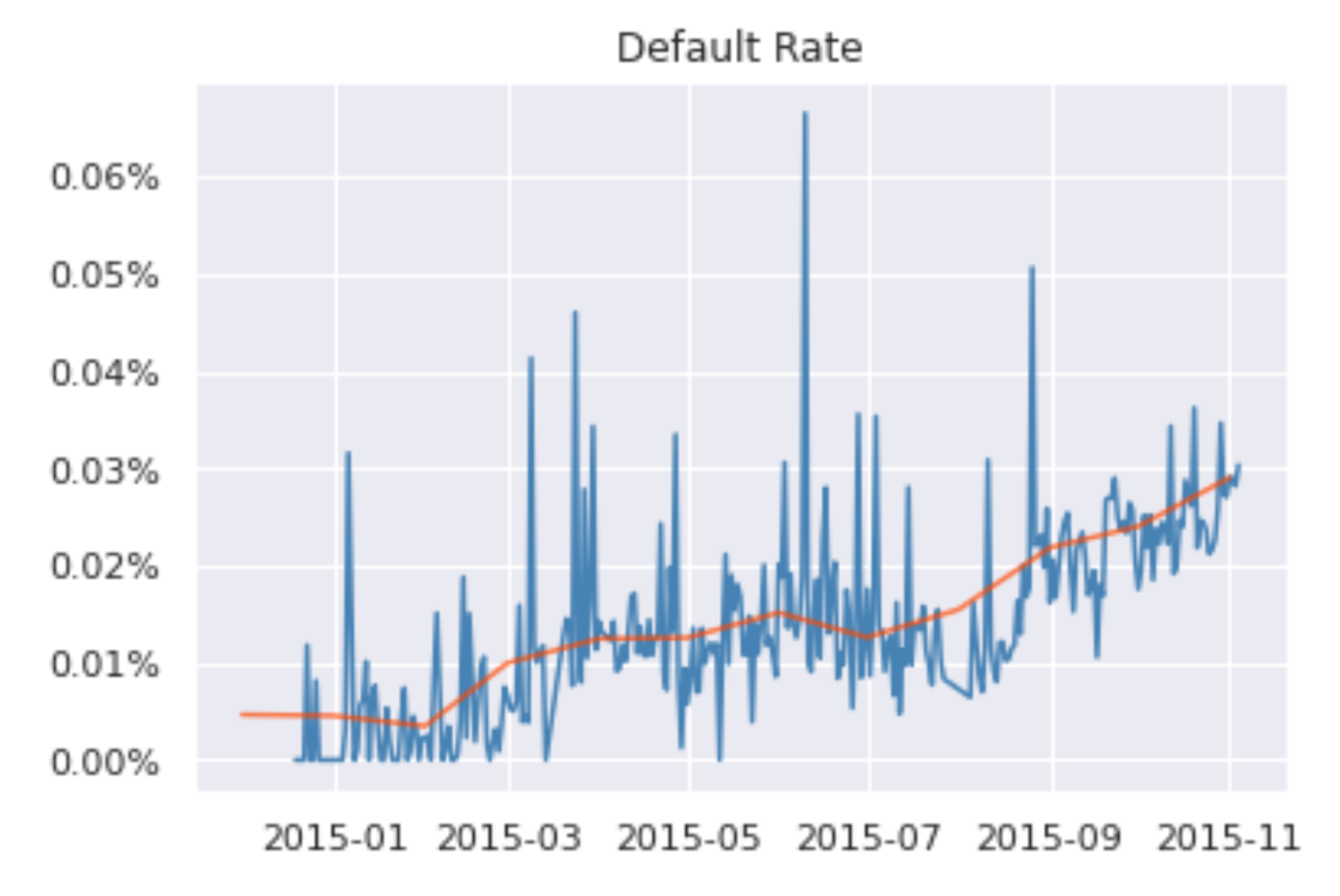
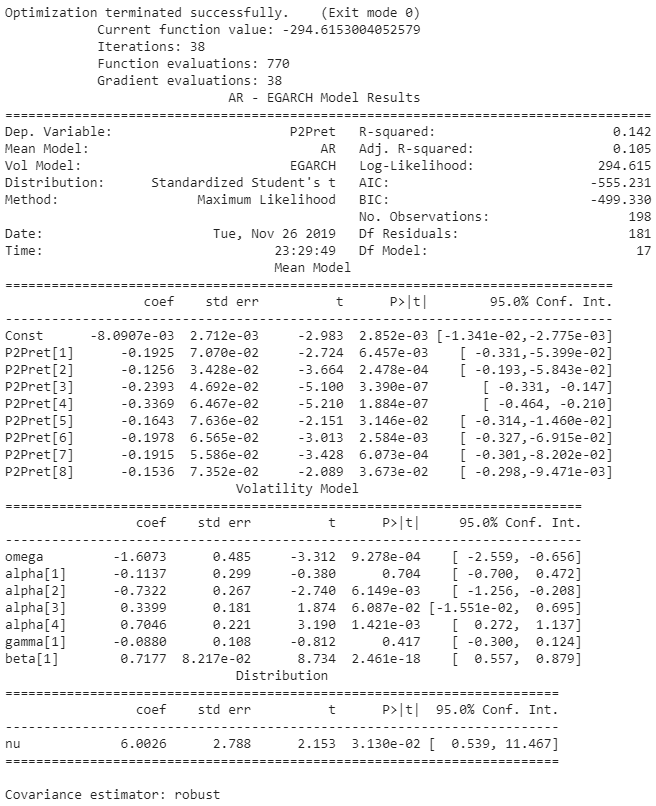


Figure . Default Rate Over Time

Regression Analysis

I tried to run ARCH (family) model on the data. I used log return, which passed the unit root test, to run the regression with 8 lags. The results are shown as below:

Table . Regression Result



The p-values are significant; however, the signs of the coefficients are all negative, which indicates that previous positive return will always have a negative impact on future returns, which is not reasonable. This may result from the data per se, as it comes through the stock market fall and the economy is going downwards. Hence, there is not fruitful in sticking to econometric analysis with this given data.

Part II: Machine Learning Analysis

Problem Statement

In machine learning, imbalanced learning a very important and common area of research. It focuses on learning the patterns of data from a dataset with an imbalance distribution, which is often the case in practice. In this project, specifically, the focus is the difficulty of binary classification in extremely imbalanced situations. An alias for imbalanced learning is positive-unlabeled (PU) learning. Given a set of examples of an particular class P (called the positive class) and a set of unlabeled examples U, which contains both class P and non-class P (called the negative class) instances, the goal is to build a binary classifier to classify the test set T into two classes, positive and negative. In conclusion, PU learning is an approach to learn binary classifiers when only positive and unlabeled instances (samples) are available.

A dataset is called “class-imbalanced” if the number of samples in the dataset of a classification task differs greatly. Take the task in this project as an example, the dataset used is the user information from an online lending platform. The target of the task is to predict the probability of default of a user by learning from their characteristics. However, few users tend to default, as basic risk control will prevent certain users from borrowing money, otherwise the platform is going to lose money for sure. This, therefore, leads to a huge gap of the number of positive/negative samples in the training/fitting dataset. The same situations are prevalent in many other practical scenarios, such as insurance frauds (normal/fraudulent claim), medical diagnosis (normal/ill), the recommendation prediction (accept/not accept), and etc. (Gamberger et al.; Graepel et al.; Sun et al.) To note, even though the minor class has fewer samples and lower quality of data, it often carries important information. Hence, the model’s capability of correctly classifying minor samples are paid more attention.

Notations

There exist only two class in an imbalanced binary classification – the one with fewer samples (called the minor/minority class) and the one with more samples (called the major/majority class). We denote the set of all training data samples as , and every sample as . In binary classification, a sample belongs to the positive class (minority class) if , and is in the negative class (majority class) if . Hence, we can define the minority class as:

And the majority class as:

Where we have:

For highly imbalanced dataset, we have . In order to quantify the imbalance of different datasets, it is a convention to use Imbalance Ratio (IR), the ratio of the number of the majority samples to the number of the minority samples.

A classification task may be slightly skewed, or alternatively, highly skewed where there are hundreds of samples in one class and only tens of samples in another class for a training dataset.

* Slight imbalance: An imbalanced classification task where the distribution of examples is uneven by a tiny amount in the training dataset (e.g. 4:6).
* Severe imbalance: An imbalanced classification task where the distribution of examples is uneven by a great amount in the training dataset (e.g. 1:100 or more).

A slight imbalance is not a concern in most cases. However, a severe imbalance would be challenging for the model and requires special techniques that tailor the problem. In real-life applications, unfortunately, the latter is often the case, such as fraud detection whose imbalance ratio could range from 1:1000 to even 1:10000 (Krawczyk). In this project, the imbalance ratio is around 1:2000.

Model Criteria

The accuracy of classification is no doubt the most common and useful criterium for most classification task. Accuracy is calculated simply by dividing the accurate prediction by the total prediction. However, when it comes to imbalanced learning, accuracy is hardly satisfying. Consider the following simple scenario: a tiny dataset with 100 samples has 5 positive samples and 95 negative ones, i.e., . Say now we have a classifier that classifies every sample as negative (majority class), this classifier has an accuracy rate as high as 95%, a high yet unreasonable score. Why could a classifier that does not distinguish anything achieve such a high score? The intrinsic nature of accuracy tends to encourage the classifier to label a sample as the class that accounts for the majority of the dataset, which producs a high accuracy but performs badly in the minority class (our focus), preventing itself from being used.

Under such circumstances, confusion matrix comes into use (*Confusion Matrix - Wikiwand*). A confusion matrix is based on True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Table 2. Confusion Matrix for Binary Classification

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | Positive | Negative |
| Predict | Positive | TP | FP |
| Negative | FN | TN |

To clarify, TP is equivalent to hit, TN equivalent to correct rejection, FP equivalent with false alarm (Type I error), and FN equivalent with miss (Type II error). Below is a table of commonly used ratios and their calculations. The ratios mainly used in the paper are precision and recall.

Table 3. Common Rate and Calculation

|  |  |
| --- | --- |
| Term | Calculation |
| Precision, or positive predictive valve (PPV) |  |
| Recall, sensitivity, or true positive rate (TPR) |  |
| F1-score |  |
| Accuracy |  |

Solution Overview

Normal machine learning algorithms assume a similar number of samples of different classes. Therefore, the uneven distribution of samples makes it difficult to apply normal algorithms on unbalanced datasets: the implicit optimization goal behind the design of these learning algorithms is the classification accuracy on the data set, which, as discussed before, will cause the learning algorithm to have bias/preference towards the majority category with more samples. Most imbalanced learning algorithms are proposed to solve the preference for the majority class.

Faced with the problem of imbalanced dataset when applying machine learning models in real life applications, scholars have developed a number of methods to improve traditional algorithms or learning processes to fit the characteristics of imbalance, in order to realize a comparably higher recall and precision rate. For now, popular methods can be categorized into data-level methods (re-sampling), algorithm-level methods (cost-sensitive models), and ensembles.

Data-level methods

Data-level methods are the earliest developed, most influential, and most widely used methods in the field of unbalanced learning. They are be called resampling methods. This type of method focuses on modifying the training dataset so that normal machine learning algorithms can be directly and effectively trained on the modified dataset. Depending on the implementation, data-level methods can be further classified as:

1. Undersampling: the method of removing samples from the majority class, examples include Random Undersampler (RUS) and Edited Nearest Neighbor (ENN) (Wilson).
2. Oversampling: the method of generating samples from the minority class, examples include SMOTE and ADASYN (Chawla et al.; He et al.).
3. Combined methods: a hybrid of undersampling and oversampling.

Algorithm-level methods

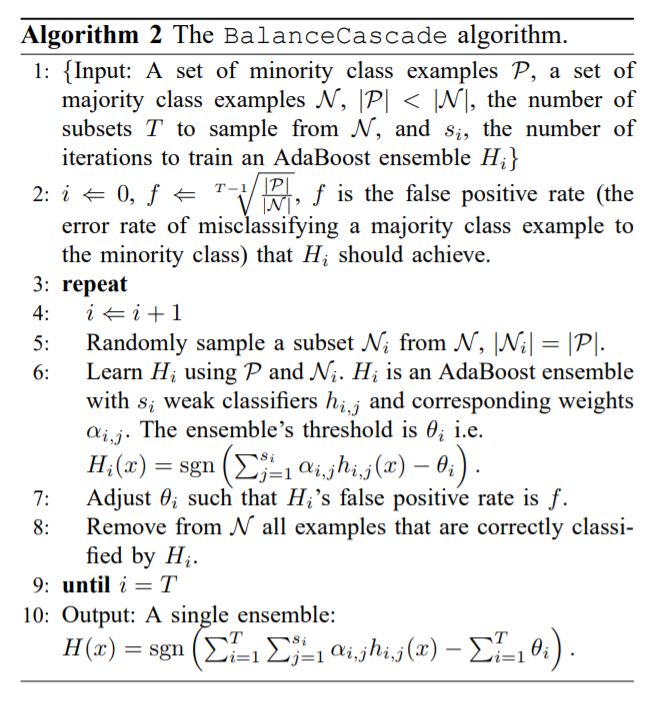
These methods modify the bias over majority class of traditional machine learning algorithms. The most popular method is cost-sensitive learning. This method gives a higher cost/punishment for misclassifying a true minority class, and a smaller cost vice versa. For example, traditionally accuracy is the only index that tells the performance of the model. By replacing accuracy with other criteria, the importance/weight of the minority class is lifted when training a classifier using cost-sensitive learning, reducing the bias towards the majority class.

Ensembles

Ensembles combine a data-level or an algorithm-level approach with ensemble learning. Because of its excellent performance in imbalanced classification tasks, ensembles are becoming more and more popular in practical applications. Most of them are based on a specific ensemble learning algorithm such as adaboost (Freund and Schapire) and embed an additional unbalanced learning method.

The biggest strength of using ensembles is its results. So far, the most effective methods of solving imbalance learning tasks are a competition of different ensembles. Besides, it is able to adjust dynamically according to the feedback during iterations. Some ensembles can resample dynamically. For example, some majority class samples are already well-classified by the current classifier, and therefore can be discarded in every iteration, as they no longer have useful information.

In this paper, I will use the BalanceCascade (Liu et al.), an ensemble method. This method uses AdaBoost (*AdaBoost - Wikiwand*) as the base classifier. The principle is that upon every iteration, it uses the same amount of major and minor training data to make prediction. By modifying the threshold to control the FP (false positive rate), it removes the data that it makes correct predictions to reduce the majority class samples.



Result

The model conducts experiments by using methods with and without addressing imbalance and compare the results of each.

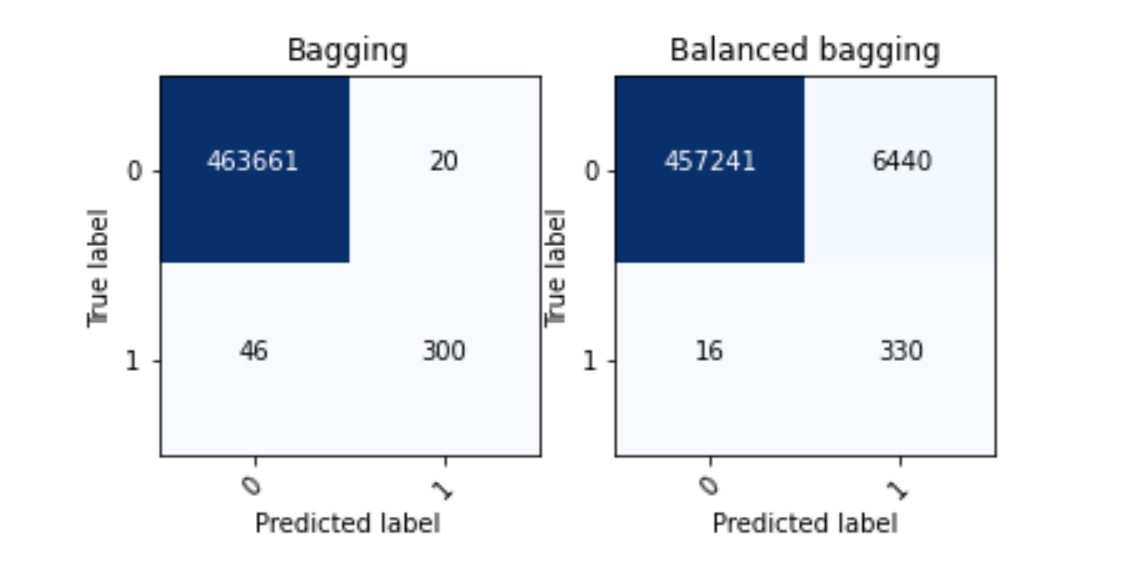


Figure 8. Result of Bagging

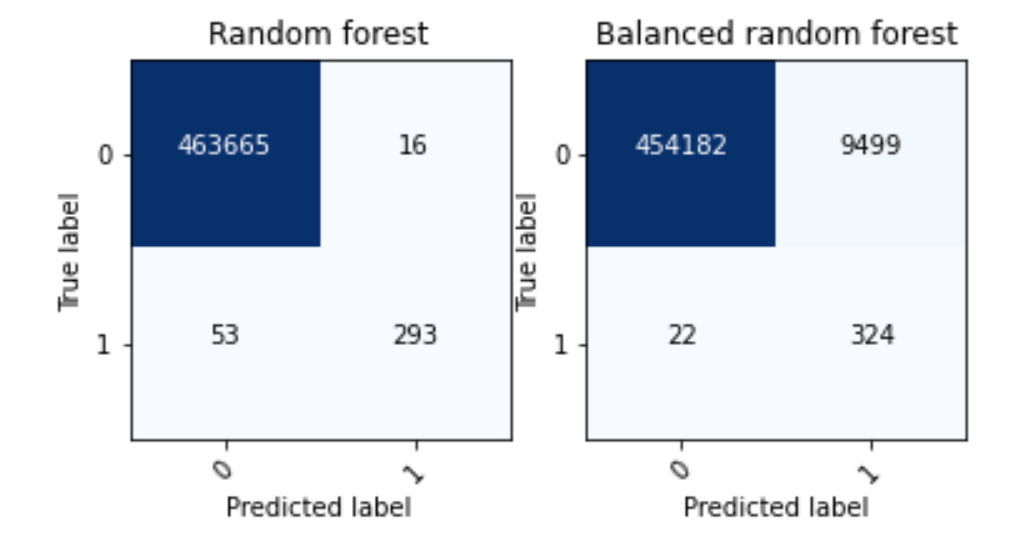


Figure 9. Result of Random Forest

Here we use a test dataset that includes a sample 460 thousand users. We can see that even though the simple model achieves a precision of 85%, the balanced model increases it by 10% to 95%. The cost is that there are more people who are not default users are labeled as default. However, compared with the large base, it is only an 1% increase, which is tolerable. Moreover, the truth that they are not default users does not say that they will not be – maybe they are highly risky people. Therefore, the modified model has done a great job so far.

Future Work

In the future, I plan to refine the model by running cross validation and modify threshold to fine tune the parameters and decrease FP. Meanwhile, I will also do more research on the trade-off between precision and recall rate, to make the modified model even stronger.

The training part of the model uses the implementation of a Python module called imbalanced-learn (Lemaıtre and Nogueira). It a easy-to-use library that incorporates various imbalanced learning methods. However, with all due respect, some of the implementations still have space for improvements. Hence, if time permitted, I will try to write my implementation of imbalanced learning approaches, and based on that, make modifications to achieve better results.

References

*AdaBoost - Wikiwand*. https://www.wikiwand.com/en/AdaBoost. Accessed 30 Mar. 2020.

Chawla, Nitesh V., et al. *SMOTE: Synthetic Minority over-Sampling Technique*. AI Access Foundation, 1 June 2002.

*Confusion Matrix - Wikiwand*. https://www.wikiwand.com/en/Confusion\_matrix. Accessed 2 May 2020.

Freund, Yoav, and Robert E. Schapire. “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.” *Journal of Computer and System Sciences*, vol. 55, no. 1, Aug. 1997, pp. 119–39. *ScienceDirect*, doi:10.1006/jcss.1997.1504.

Gamberger, Dragan, et al. “Experiments with Noise Filtering in a Medical Domain.” *Proceedings of the Sixteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., 1999, pp. 143–151.

Graepel, Thore, et al. “Web-Scale Bayesian Click-through Rate Prediction for Sponsored Search Advertising in Microsoft’s Bing Search Engine.” *Proceedings of the 27th International Conference on International Conference on Machine Learning*, Omnipress, 2010, pp. 13–20.

He, Haibo, et al. “ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning.” *In: Ieee International Joint Conference on Neural Networks (Ieee World Congress on Computational In℡ligence), Ijcnn 2008*, 2008, pp. 1322–1328.

He, Qizhi, and Mingsheng Peng. “基于互联网金融的网贷利率特征研究.” 金融研究, no. 10, 2016, pp. 95–110.

Krawczyk, Bartosz. “Learning from Imbalanced Data: Open Challenges and Future Directions.” *Progress in Artificial Intelligence*, vol. 5, no. 4, Nov. 2016, pp. 221–32. *Springer Link*, doi:10.1007/s13748-016-0094-0.

Lemaıtre, Guillaume, and Fernando Nogueira. *Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning*. p. 5.

Liu, Xu-Ying, et al. *Exploratory Undersampling for Class-Imbalance Learning*. IEEE Press, 1 Apr. 2009. *ACM Digital Library*, https://doi.org/10.1109/TSMCB.2008.2007853.

“P2P Online Lending Platform 2019 Annual Report.” 未央网. *www.weiyangx.com*, https://www.weiyangx.com/347574.html. Accessed 24 Mar. 2020.

Sun, Yanmin, et al. “Classification of Imbalanced Data: A Review.” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 04, World Scientific Publishing Co., June 2009, pp. 687–719. *worldscientific.com (Atypon)*, doi:10.1142/S0218001409007326.

Wilson, Dennis L. “Asymptotic Properties of Nearest Neighbor Rules Using Edited Data.” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-2, no. 3, July 1972, pp. 408–21. *IEEE Xplore*, doi:10.1109/TSMC.1972.4309137.