symmetry

MDPI

Article

Accessing Information Asymmetry in Peer-to-Peer Lending by Default Prediction from Investors' Perspective

Xinyuan Wei [1,*], Bo Yu [1] and Yao Liu [2]

[1] School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China; yubo@dlut.edu.cn
[2] Simon Business School, University of Rochester, Rochester, NY 14627, USA; yao.liu@garp.com
* Correspondence: xinyuanwei@mail.dlut.edu.cn

Received: 26 April 2020; Accepted: 29 May 2020; Published: 3 June 2020

Abstract: Recent a few years have witnessed the rapid expansion of the peer-to-peer lending marketplace. As a new field of investment and a novel channel of financing, it has drawn extensive attention throughout the world. Many investors have shown great enthusiasm for this field. However, investors are at the disadvantage of information asymmetry, which is a key issue in this marketplace that is unavoidable and can lead to moral hazard or adverse selection. In this paper, we propose an $L_{1/2}$-regularized weighted logistic regression model for default prediction of peer-to-peer lending loans from investors' perspective, which can reduce the impact of information asymmetry in the process of loan decision. Rather than solely focus on the accuracy of the prediction, we take into consideration the different risk preferences of different investors. We try to find a trade-off between the risk of losing principal and that of losing potential investment opportunities on the basis of investors' risk preferences. Meanwhile, due to the nature of peer-to-peer lending loans, we add an $L_{1/2}$-regularization term to reduce the chance of overfitting. Xu's algorithm for $L_{1/2}$-regularization problems is applied to solve our model. We perform training, in-sample test, and out-of-sample test with data from LendingClub. Numerical experiments demonstrate that regularization could enhance out-of-sample the area under the Precision–Recall curve (AUPRC). By applying the proposed model, the risk-averse investors could apply a higher penalty factor to lower the risk of losing principal at the cost of the loss of some potential investment opportunities according to their own risk preferences. This model can help investors reduce the impact of information asymmetry to a great extent.

Keywords: default prediction; information asymmetry; $L_{1/2}$-regularization; weighted logistic regression; peer-to-peer lending

1. Introduction

Peer-to-peer lending (also known as people-to-people lending, person-to-person lending, or social lending), often shorted as P2PL, a form of crowdfunding, is an online practice of individuals or businesses lending money to other individuals or businesses without going through a traditional financial intermediary. A classical P2PL model involves three basic elements: investors (supply), borrowers (demand), and a platform. In the modern financial market, investors have a variety of choices, such as stocks, bonds, futures. However, P2PL enables small investments as low as $25, which may have little chance of investment elsewhere. Meanwhile, it would help investors diversify their traditional portfolios. Additionally, interest rates offered by P2PL are usually more competitive than those of traditional banks while it can build connections between borrowers and investors faster and cheaper than any bank. Compared to stock markets, P2PL investments enjoy lower volatility and correlation. These merits make it a good alternative to traditional investments.

However, investors in this marketplace should be extremely cautious since its special risk characteristics. Loan applicants are individuals with all kinds of uncertainty. Default is more likely to happen than bonds or T-bills. Information asymmetry is a key issue in this marketplace, which can result in moral hazard or adverse selection [1]. As it comes to the loan decision, investors are at a disadvantage to the borrower, where the borrower has near-complete information while the investors can only access the information provided by the platform. Though P2PL platforms seek to reduce the impact of information asymmetry via many mechanisms, investors should also take information asymmetry into consideration in loan decision. From investors' perspective, an effective default prediction would help to protect their profits and principle in such a marketplace. P2PL platforms usually provide a mass of information, thought not as much as that possessed by the borrower, which will help investors in loan decision making.

In the next section, we will introduce the peer-to-peer lending marketplace in detail.

2. Theoretical Background

2.1. Development of Peer-to-Peer Lending in Marketplace

As a novel financial model, P2PL has attracted public attention over the past decade when many P2PL companies came into being across the world.

The first company to offer peer-to-peer loans in the world, ZOPA, was founded in Britain in 2005. The name, ZOPA, which stands for "zone of possible agreement", is a negotiating term that identifying the bounds within which agreement can be reached between the two parties [2]. Prosper Marketplace, the first P2PL company in the United States, was also founded in 2005. It began operations in February 2006 and was the only P2PL company in the United States until May 2007, when LendingClub was founded. In the beginning, Prosper issued loans to anyone who had the interest to get a loan, which caused most of its investors to get negative returns. At that time, Prosper offered only unsecured consumer loans but not small- and medium-sized enterprise (SME) loans. In 2008, Prosper was temporarily shut down because of scrutiny by the Securities and Exchange Commission (SEC). SEC issued a formal cease-and-desist letter to explain that Prosper should be considered as a seller of securities and should be regulated by the SEC [3].

LendingClub was first introduced as a Facebook application. With rapid growth, it emerged as a standalone website within a couple of months. It was the first P2PL company that registered its offerings as securities with SEC. It offers loans from $1000 to $35,000 for individuals and from $15,000 to $300,000 for SME. Currently, LendingClub is the largest P2PL platform in the world.

In 2007, TrustBuddy, the first P2PL company in Sweden, began operations. Now it is a peer-to-peer group that operates in five European countries under three different brand names (Geldvoorelkaar, Crowdfunding Society and TrustBuddy).

The first P2PL company in China was also set up in the year 2007, named "Paipaidai". This marketplace has undergone extremely rapid growth in the past few years. In 2015, the national P2P net loans turnover has increased 258.62% compared to the year 2014 and reached RMB 1180.6 billion and 3844 platforms reported to be operating [4].

Funding Circle, a P2PL platform founded in the UK in August 2010, entered the US in October 2013. It only processes SME loans and operates in the US, UK, Germany, and the Netherlands.

Upstart, founded in April 2012 in San Carlos, California, by a group of ex-Googlers, was first launched with an Income Share Agreement (ISA) product that enabled individuals to raise money by contracting to share a portion of future income. Later, it pivoted away toward the personal loan marketplace. Upstart operates differently in many ways from other P2PL platforms. The firm specifies its target niche as young professionals. It applies unique grading criteria taking into consideration not only Fair Issac Credit Organization (FICO) scores but also educational background information and employs a so far remarkably accurate modeling system at predicting future defaults and returns. This helps the firm enjoy the lowest default rates across the P2PL industry up to 2017.

Some other countries also opened up P2PL industry in recent years, such as Australia, India, Israel, Canada, and Brazil.

2.2. Literature Review

Although P2PL is a relatively young field of research, it has been extensively studied in the past decade. Since the first P2PL platform ZOPA launched, research on this new lending pattern gains increasing attention. Wang et al. [5] provide an overview of the concepts and discussed some different P2PL marketplace models in detail. Prosper and LendingClub gave great impetus to research on P2PL by giving full public access to their data. Traditional research work on P2PL mainly focused on funding success, that is, looking for the features with which loan applicants are more likely to succeed, such as [6,7]. Among a variety of research topics on P2PL, default prediction has always been in the spotlight since its significance for borrowers. Ajay et al. [8] propose a credit scoring model to perform default prediction based on artificial neural networks. They are also aiming to reduce the risk of investment failure. The numerical results show a 64.47% of the non-default loans and 74.75% of the default loans are correctly classified for training data while 62.70% of the non-default loans and 74.38% of the default loans are correctly classified for testing data. Jiang et al. [9] apply a tex analysis method and latent Dirichlet allocation (LDA) model to extract soft information from text to be combined with hard information. Then they present a prediction model based on a two-stage feature selection method. Kim and Cho [10] consider an ensemble semi-supervised learning method taking into account both labeled data and unlabeled data.

Other research mainly includes investment strategy designation, the role of P2PL in financial market, information asymmetry, interest rate, etc., to name a few [11–15].

2.3. Peer-to-Peer Lending Process

For a potential borrower, the first step is to submit an application to a P2PL platform, which usually contains the information about the borrower and the loan he would like to apply for, such as loan amount, annual income, and Social Security Number (SSN).

After receiving the application, the platform will access the status of the potential borrower with its own system taking into account information provided by the applicant and also the information obtained through the applicant's SSN, such as Fair Issac Credit Organization (FICO) score, debt-to-income (DTI) ratio, and other credit information. Based on this information, the platform decides whether to approve the loan. This process is usually called loan application processing. Different platforms may differ in loan application processing scheme and also in the way to set the interest rate.

Once a loan is approved by the platform, detailed information about the loan and the applicant will go public online. Potential investors have a period of time to review the loan information and make the decision to invest or not. A loan is issued if it collects enough funding within this period of time; otherwise, the loan is dismissed and the money collected will go back to investors' accounts.

After the loan is issued, the borrower gets the money collected and makes monthly payment to repay. The platform charges a scheduled rate of fee for service.

Although platforms tried to provide qualified loans with complex loan application processing systems, investors may get negative returns at the maturity of the loan due to the investment risks involved in P2PL.

2.4. Investment Risk of Peer-to-Peer Lending

Investment in P2PL may face many types of risks, just as other financial instruments do, including but not limited to: default risk, bankruptcy risk, regulatory risk, interest rate risk, prepayment risk, and liquidity risk.

The main risk in P2PL is default risk, which related to the loans selected to invest, i.e., investors' investment strategies will affect the default risk exposure of a portfolio to a great extent. Other types

of risks may not have as much effect as default risk since the risk events may be unlikely to happen or measured in the sense of opportunity costs. We would like to introduce several main risks to investors below.

2.4.1. Default Risk

Default risk is the chance that borrowers may be unable to repay their loans entirely or partially, and it is the main risk that investors in P2PL will encounter. Many works have investigated into default prediction, see [16], including default prediction in P2PL [17–19]. However, these works depend on meta-level phone usage data, which is not available for general investors.

2.4.2. Bankruptcy Risk of P2PL Platform

Investors of P2PL may face the risk that platforms shut down, especially when the P2PL industry goes crazy. For example, in 2011, Quakle, a UK P2PL company closed down with a nearly 100% default rate due to the unsuccessful attempt to measure borrowers' creditworthiness. This type of risk is closely related to default risk. We could go further and say bankruptcy risk of P2PL platforms mainly caused by borrows default.

However, this type of risk is fairly low in the current stable economic environment. With the improving regulatory enforcement, choosing a legal compliance P2PL platform could help to reduce the bankruptcy risk of the platform to a negligible level.

2.4.3. Regulatory Risk

Regulatory risk is the risk that a change in regulations or laws which will materially impact the whole industry. Generally, events which involve regulatory risk occur in the early years of market establishment, when the market is premature or when notable events happen. LendingClub temporarily shut down lending operations from April 2008 to October 2008 and Prosper did not offer investment opportunity from October 2008 to July 2009. Both platforms were preparing to file the registration statement with the SEC [20]. In China, at least 246 P2PL platforms were shut down during the first half of 2016 since tightening of regulation according to a report by cnr.cn.

However, most of the time, regulatory risk is unpredictable and uncontrollable. Fortunately, it is unlikely to happen when the market is in normal operation.

2.4.4. Interest Rate Risk

Interest rate risk is the risk that arises for fixed income securities owners from interest rates fluctuation. As reported by SEC, all bonds are subject to interest rate risk, even if they are insured or government guaranteed. This type of risk is mainly affected by the overall economic climate and maturity of the security. That is, securities in the same market and with the same maturity face similar interest rate risk. Loans on one platform in P2PL are of this kind of situation.

2.4.5. Prepayment Risk

Platforms usually allow extra payments and full prepayment. These payments could be made any time and would be applied directly to the borrower's principal balance. It would decrease the total cost of the loan by reducing the principal balance and the total interest that borrowers pay on this amount. That is, for investors, prepayment would reduce the return lower than a prospective return.

2.4.6. Liquidity Risk

Investors of P2PL would also face liquidity risk, which is the risk that stems from the lack of marketability. In the case of LendingClub, investors should be prepared to hold any note purchased through to its maturity. Even though there is a secondary market, Folio Investing, there is no guarantee that investors will find buyers for their notes. This type of risk is common in most bond markets.

Due to the risk characteristics involved, default events happen from time to time. This makes default prediction necessary for investors, especially for this marketplace has a high level of information asymmetry. From historical statistics, we can see that default loans are relatively few compared to loans successfully repaid. Taken default prediction as a binary classification problem would confront the problem of class imbalance. Meanwhile, overfitting is another problem since there are too many features in P2PL data, especially considering the introduction of dummy variables, while simply deleting some of them may cause loss of information. Additionally, different investors may have different risk preferences, which makes traditional classification models impracticable for every investor.

In this paper, from the investors' perspective, we develop an $L_{1/2}$-regularized weighted logistic regression model for default prediction of P2PL loans. A penalty factor on the negative class is applied to deal with class imbalance. Additionally, by adjusting this parameter, investors can weigh the risk of losing principal and that of potential investment opportunities according to their own risk preferences. The introduction of $L_{1/2}$ regularizer help to reduce the chance of overfitting. We also give out a proof of the convergence of Algorithm 1 for this model. Finally, we test the performance of $L_{1/2}$-regularized weighted logistic regression model by applying it to the data from LendingClub.

Algorithm 1 Xu's Algorithm

Set the initial value $\tilde{\boldsymbol{\beta}}^0 = [1, 1, \ldots, 1]^\top \in \mathbf{R}^{m+1}$ and the tolerance ϵ, where $\epsilon > 0$ is a small value much larger than machine precision. Let $t = 0$.

repeat

Solve

$$\tilde{\boldsymbol{\beta}}^{t+1} = \arg\min\{w(\tilde{\boldsymbol{\beta}}) + \frac{\lambda}{2}\sum_{i=1}^{m}\frac{1}{\sqrt{|\beta_i^t|}}|\beta_i|\}^1,$$

until $\|\tilde{\boldsymbol{\beta}}^{t+1} - \tilde{\boldsymbol{\beta}}^t\|_\infty < \epsilon$.

The rest of this paper is organized as follows. In Section 3, we establish the $L_{1/2}$-regularized weighted logistic regression model and explain its application in default prediction. We apply Algorithm 1 to solve this model, and we give out a proof of the convergence result. In Section 4, we explain the performance measure in use. We carry out numerical experiments with the data from LendingClub to test the performance in Section 5. Finally, we come to a conclusion in Section 6.

3. Default Prediction by $L_{1/2}$ Regularized Weighted Logistic Regression

Throughout the duration of a loan, there would be several types of loan statuses. Here, we only focus on the statuses possibly at the expiration.

For LendingClub, loans may take one of the following statuses (For more details of loan statuses on LendingClub, see https://help.lendingclub.com/hc/en-us/articles/215488038-What-do-the-different-Note-statuses-mean-) at its predetermined maturity date.

- Fully Paid: The loan has been fully repaid, either at the expiration of the 36- or 60-month term, or as a result of a prepayment.
- Current: The loan is up to date on all outstanding payment.
- In Grace Period: There will be a 15-day grace period if the loan past due.
- Late (16–30): The loan has not been current for 16 to 30 days.
- Late (31–120): The loan has not been current for 31 to 120 days.
- Default: The loan has past due for more than 121 days.
- Charged Off: The loan for which there is no longer a reasonable expectation of further payments. Upon Charge Off, the remaining principal balance of the note is deducted from the account balance.

Usually, the platform has a complicated loan applications processing scheme to determine whether to issue or reject a loan application. It helps to distinguish qualified loan applications from unqualified ones to a great extent. For example, up to the first quarter of 2019, LendingClub has issued about 2 million loans, while more than 30 million loans have been declined which account for 93.78% of total loan applications. However, among the issued loans, only about 0.96 million loans have been fully paid, and about 1.1 million are with the status "Current", which means the loan is up to date on all outstanding payment. There are still about 0.28 million loans not likely to be paid back with statuses "In Grace Period", "Late (16–30)", "Late (31–120)", "Default", or "Charged Off", which would lead to significant capital loss to investors. Detailed loan status statistics of the loans issued up to the first quarter of 2019 are shown in Table 1 (Data are drawn from LendingClub, https://www.lendingclub.com).

Table 1. Loan status statistics up to the 1st quarter of 2019.

Date	B	Loan Status						
		Fully Paid	Current	In Grace Period	Late (16–30)	Late (31–120)	Default	Charged Off
2007–2011	42,536	36,104	0	0	0	0	0	6431
2012–2013	188,181	156,882	1573	38	26	63	3	29,596
2014	235,629	177,018	17,130	344	166	549	25	40,397
2015	421,095	269,699	74,267	1514	773	549	148	71,761
2016Q1	133,887	57,828	52,660	863	522	1688	68	20,258
2016Q2	97,854	38,441	42,841	899	345	1410	66	13,852
2016Q3	99,120	37,507	45,454	760	497	1696	81	13,125
2016Q4	103,546	35,528	53,924	948	543	1948	93	10,562
2017Q1	96,779	28,528	57,267	827	494	1888	78	7697
2017Q2	105,451	25,806	68,711	1113	606	2350	136	6729
2017Q3	122,701	23,521	88,386	1387	820	2933	127	5527
2017Q4	118,648	17,082	94,116	1255	678	2346	117	3054
2018Q1	107,864	10,004	93,464	977	581	1707	73	1058
2018Q2	130,772	6601	120,933	1146	546	1292	26	228
2018Q3	128,198	17,275	103,708	923	483	2425	3	3377
2018Q4	128,416	11,786	111,679	730	525	2146	2	1544
2019Q1	115,679	6184	107,128	549	355	1109	0	350

We train the model with loans that already past the predetermined maturity, where "Current" means the borrower must have missed or been late for at least one payment. Throughout this paper, we take "Fully Paid" as one category and all the others as the other category, named "Not Fully Paid". As shown in Table 1, the datasets are highly imbalanced. Therefore, the default prediction turns into a binary classification problem with class imbalance. In this binary classification, we take the status of loans as the target variable, where 1 denotes Fully Paid and 0 denotes Not Fully Paid; while, the independent variables are chosen from features of loans accessible to investors. We will discuss the features in detail later in Section 5.1.

Notation: Suppose we have a sample of size n,

$$S = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \cdots, (\mathbf{x_n}, y_n)\} \subseteq \mathbf{R}^n \times \mathbf{Y},$$

where $\mathbf{x_i} = (x_{i1}, x_{i2}, \cdots, x_{im})^\top$ and $n, m \in \mathbf{N}^+$. Here, $x_{ij} \in \mathbf{R}$ represents the jth feather of the ith loan and y_i is the loan status of the ith loan taken from $\mathbf{Y} = \{0, 1\}$, where 0 represents Not Fully Paid (negative class) and 1 represents Fully Paid (positive class). Without loss of generality, we assume any two loans are independent. That is, if one borrower defaults, it is not likely to affect the probability of a default event of any other borrower.

3.1. Standard Logistic Regression

Logistic regression is a machine learning algorithm borrowed from statistics. It is an important topic in both fields. Since y is the label, it is an indicator variable taking value from $\mathbf{Y} = \{0, 1\}$. Obviously, $\text{Prob}(y = 1) = \mathbf{E}[y]$. Then, the conditional probability is the conditional expectation of the

indicator, i.e., $\text{Prob}(y = 1|X = \mathbf{x}) = \mathbf{E}[y|X = \mathbf{x}]$. Denote the p-value involved with some parameter $\boldsymbol{\beta} \in \mathbf{R}^m$ as $p(\mathbf{x}; \tilde{\boldsymbol{\beta}}) = \text{Prob}(y = 1|X = \mathbf{x})$, where $\tilde{\boldsymbol{\beta}} = [\beta_0, \boldsymbol{\beta}^\top]^\top = (\beta_0, \beta_1, \cdots, \beta_m)^\top \in \mathbf{R}^{m+1}$.

From the independence of $\mathbf{x_i}$, we have

$$\prod_{i=1}^{n} \text{Prob}(y = y_i|X = \mathbf{x_i}) = \prod_{i=1}^{n} p(\mathbf{x_i}; \tilde{\boldsymbol{\beta}})^{y_i} (1 - p(\mathbf{x_i}; \tilde{\boldsymbol{\beta}}))^{1-y_i}. \tag{1}$$

In the standard logistic regression model, the conditional probability distribution of the label y given the feature vector $\mathbf{x}$ can be formed as

$$\text{Prob}(y = 1|\mathbf{x}) = g(\tilde{\boldsymbol{\beta}}^\top \tilde{\mathbf{x}}) = \frac{1}{1 + \exp(-\tilde{\boldsymbol{\beta}}^\top \tilde{\mathbf{x}})}, \tag{2}$$

and

$$\begin{aligned} \text{Prob}(y = 0|\mathbf{x}) &= 1 - g(\tilde{\boldsymbol{\beta}}^\top \tilde{\mathbf{x}}) \\ &= 1 - \frac{1}{1 + \exp(-\tilde{\boldsymbol{\beta}}^\top \tilde{\mathbf{x}})} \\ &= \frac{1}{1 + \exp(\tilde{\boldsymbol{\beta}}^\top \tilde{\mathbf{x}})}, \end{aligned} \tag{3}$$

$$\begin{aligned} \text{Prob}(y = 0|\mathbf{x}) &= 1 - g(\tilde{\boldsymbol{\beta}}^\top \tilde{\mathbf{x}}) \\ &= 1 - \frac{1}{1 + \exp(-\tilde{\boldsymbol{\beta}}^\top \tilde{\mathbf{x}})} \\ &= \frac{1}{1 + \exp(\tilde{\boldsymbol{\beta}}^\top \tilde{\mathbf{x}})}, \end{aligned} \tag{4}$$

where $\tilde{\mathbf{x}} = [1, \mathbf{x}^\top]^\top \in \mathbf{R}^{m+1}$. Here, $g(z)$ is the logistic function (also known as sigmoid function) defined as

$$g(z) = \frac{1}{1 + \exp(-z)}. \tag{5}$$

The standard logistic regression model can be built by minimizing the negative log-likelihood (NLL) $f(\tilde{\boldsymbol{\beta}})$,

$$\min_{\tilde{\boldsymbol{\beta}} \in \mathbf{R^{m+1}}} \quad f(\tilde{\boldsymbol{\beta}}) = -\sum_{i=1}^{n} [y_i \log(p(\mathbf{x_i}; \tilde{\boldsymbol{\beta}})) + (1 - y_i) \log(1 - p(\mathbf{x_i}; \tilde{\boldsymbol{\beta}}))]. \tag{6}$$

3.2. Weighted Logistic Regression

In loan default prediction, the Type I error (also known as False Positive), which happens when a classifier incorrectly classifies a Not Fully Paid loan as a Fully Paid loan, is more serious than the Type II error (also known as False Negative), which is the misclassification of a Fully Paid loan as a Not Fully Paid loan. That is because the Type I error will lead to real loss of capital and it is what we want to avoid at all cost; while the Type II error means loss of potential investment opportunities, which is not as dangerous as the Type I error. Thus, we are more reluctant to accept Type I errors.

Since for a given sample size, the probability of making a Type I error and that of making a Type II error cannot be reduced simultaneously, we need to judge and weight Type I and Type II errors.

Tsai, Ramiah, and Singh state that precision is a more suitable statistical measure of performance in this situation and introduce a penalty factor θ into log-likelihood [21] as,

$$\min_{\tilde{\boldsymbol{\beta}} \in \mathbf{R^{m+1}}} w(\tilde{\boldsymbol{\beta}}) = -\sum_{i=1}^{n} [y_i \log(p(\mathbf{x_i}; \tilde{\boldsymbol{\beta}})) + \theta(1 - y_i) \log(1 - p(\mathbf{x_i}; \tilde{\boldsymbol{\beta}}))], \tag{7}$$

where, $\theta > 1$ is a penalty factor (weight) on the negative class.

Obviously, for a given sample size, a high θ will decrease the probability of a Type I error, even if meanwhile it will increase the probability of a Type II error. This modification could yield higher precision at the cost of recall and prediction accuracy. Their numerical experiments on the data of LendingClub also suggest that for this problem weighted Logistic Regression outperforms LibSVM, Naïve Bayes, and Random Forest.

3.3. $L_{1/2}$ Regularized Weighted Logistic Regression

Since classical logistic regression may cause over-fitting when the sample size is not large enough compared to the dimension of features [22], i.e., $n \gg m$ does not hold. The introduction of a penalty factor on the negative class can cope with the problem of data imbalance but cannot alleviate the problem of over-fitting.

Let us consider some techniques, such as L_p regularization, which is one of several useful techniques to overcome this weakness [23] taking the form,

$$\min_{\tilde{\boldsymbol{\beta}} \in \mathbf{R^{m+1}}} \{l(\tilde{\boldsymbol{\beta}}) + \lambda \|\boldsymbol{\beta}\|_p^p\}, \tag{8}$$

where $l(\cdot)$ is a loss function; $\|\boldsymbol{\beta}\|_p = (\sum_{i=1}^m |\beta_i|^p)^{1/p}$ denotes the L_p quasi-norm. Here, $\lambda > 0$ is the regularization parameter used to weight between the loss function $l(\tilde{\boldsymbol{\beta}})$ and the regularization term $\|\boldsymbol{\beta}\|_p^p$.

Zongben et al. [22] introduce an $L_{1/2}$ regularizer since it can be solved easier than L_0 regularizer, which yields the most sparse solutions but faces the problem of combinatory optimization. Meanwhile, $L_{1/2}$ regularizer is more sparse and stable than the L_1 regularizer which often yields solutions less sparse than L_0 regularizer and is inefficient when the error follows a fat tail distribution. Moreover, Xu shows the unbiasedness and Oracle properties, and presents an iteration algorithm to solve the $L_{1/2}$ regularizer.

Hence, by taking advantages of $L_{1/2}$ regularizer, our objective is

$$\min_{\tilde{\boldsymbol{\beta}} \in \mathbf{R}^{m+1}} \{w(\tilde{\boldsymbol{\beta}}) + \lambda \|\boldsymbol{\beta}\|_{1/2}^{1/2}\}. \tag{9}$$

Zongben et al. [22] also present an iteration algorithm which transforms the solution of the $L_{1/2}$ regularizer into a series of convex weighted Lasso. Here, we apply this algorithm to solve the default prediction by a modification of the termination criterion. However, we use $\frac{1}{2}\sum_{i=1}^m \frac{1}{\sqrt{|\beta_i^t|}}|\beta_i|$ to approximate the $L_{1/2}$ regularizer instead of $\sum_{i=1}^m \frac{1}{\sqrt{|\beta_i^t|}}|\beta_i|$ which would help to correct the proof in their work.

In order to avoid the error of $\frac{1}{0}$, $\frac{1}{\sqrt{|\beta_i^t|}}$ has been replaced by $\frac{1}{\sqrt{|\beta_i^t|}+\sigma}$, where $\sigma \geq 0$ is an arbitrary small number.

In the iterative process, some of $\beta_i^t, t \geq 1, i = 1, \cdots, m$ may become zero.

Theorem 1. *Given* $\theta \geq 0$, $\tilde{\boldsymbol{\beta}}^{t+1} = \arg\min\{h(\tilde{\boldsymbol{\beta}}) := w(\tilde{\boldsymbol{\beta}}) + \frac{\lambda}{2}\sum_{i=1}^m \frac{1}{\sqrt{|\beta_i^t|}}|\beta_i|\}$ *converges to the set of stationary points of* $h^*(\tilde{\boldsymbol{\beta}}) := w(\tilde{\boldsymbol{\beta}}) + \lambda\|\boldsymbol{\beta}\|_{1/2}^{1/2}$.

Proof. From the definition of p,

$$p(\mathbf{x}_i; \tilde{\boldsymbol{\beta}}) = \frac{1}{1 + \exp(-\tilde{\boldsymbol{\beta}}^\top \tilde{\mathbf{x}}_i)},$$

we have

$$w(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^n [\theta(1 - y_i)\tilde{\boldsymbol{\beta}}^\top \tilde{\mathbf{x}}_i + (\theta + (1 - \theta)y_i)\log(1 + \exp(-\tilde{\boldsymbol{\beta}}^\top \tilde{\mathbf{x}}_i))].$$

Then, the gradient and Hessian matrix of the function $w(\tilde{\boldsymbol{\beta}})$ are as below,

$$\nabla w(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^{n}[\theta(1-y_i)\tilde{\mathbf{x}}_i - (\theta+(1-\theta)y_i)\frac{1}{1+\exp(\tilde{\boldsymbol{\beta}}^\top\tilde{\mathbf{x}}_i)}\tilde{\mathbf{x}}_i],$$

and

$$\nabla^2 w(\tilde{\boldsymbol{\beta}}) = (\theta+(1-\theta)y_i)\frac{\exp(\tilde{\boldsymbol{\beta}}^\top\tilde{\mathbf{x}}_i)}{(1+\exp(\tilde{\boldsymbol{\beta}}^\top\tilde{\mathbf{x}}_i))^2}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top.$$

Since $\theta \geq 0$, we have $\nabla^2 w(\tilde{\boldsymbol{\beta}}) \succeq 0$. That is, $\nabla^2 w(\tilde{\boldsymbol{\beta}})$ is positive semi-definite.
Now, we define a function $\hat{w}(\cdot,\cdot)$ associated with $w(\cdot)$ as,

$$\hat{w}(\tilde{\boldsymbol{\beta}}^+,\tilde{\boldsymbol{\beta}}^-) = \sum_{i=1}^{n}[\theta(1-y_i)(\tilde{\boldsymbol{\beta}}^+-\tilde{\boldsymbol{\beta}}^-)^\top\tilde{\mathbf{x}}_i + (\theta+(1-\theta)y_i)\log(1+\exp(-(\tilde{\boldsymbol{\beta}}^+-\tilde{\boldsymbol{\beta}}^-)^\top\tilde{\mathbf{x}}_i))],$$

where, $\boldsymbol{\beta}^+ = \max(\boldsymbol{\beta},0)$ and $\boldsymbol{\beta}^- = -\min(\boldsymbol{\beta},0)$, are the positive part and negative part of $\boldsymbol{\beta}$, respectively. Obviously, it holds, $\boldsymbol{\beta} = \boldsymbol{\beta}^+ - \boldsymbol{\beta}^-$, $|\boldsymbol{\beta}| = \boldsymbol{\beta}^+ + \boldsymbol{\beta}^-$, and

$$\hat{w}(\tilde{\boldsymbol{\beta}}^+,\tilde{\boldsymbol{\beta}}^+) = w(\tilde{\boldsymbol{\beta}}). \tag{10}$$

Similarly, we have the gradient and Hessian matrix of $\hat{w}(\cdot,\cdot)$ as,

$$\nabla\hat{w}(\tilde{\boldsymbol{\beta}}^+,\tilde{\boldsymbol{\beta}}^-) = \begin{bmatrix}\nabla w(\tilde{\boldsymbol{\beta}})\\ -\nabla w(\tilde{\boldsymbol{\beta}})\end{bmatrix},$$

and

$$\nabla^2\hat{w}(\tilde{\boldsymbol{\beta}}^+,\tilde{\boldsymbol{\beta}}^-) = \begin{bmatrix}\nabla^2 w(\tilde{\boldsymbol{\beta}}) & -\nabla^2 w(\tilde{\boldsymbol{\beta}})\\ -\nabla^2 w(\tilde{\boldsymbol{\beta}}) & \nabla^2 w(\tilde{\boldsymbol{\beta}})\end{bmatrix}.$$

From the positive definiteness of $\nabla^2 w(\tilde{\boldsymbol{\beta}})$, we have $\nabla^2\hat{w}(\tilde{\boldsymbol{\beta}}^+,\tilde{\boldsymbol{\beta}}^-) \succeq 0$. Thus, the function $\hat{w}(\cdot,\cdot)$ is convex in $(\tilde{\boldsymbol{\beta}}^+,\tilde{\boldsymbol{\beta}}^+)$, i.e.,

$$\hat{w}(\tilde{\boldsymbol{\beta}}^+,\tilde{\boldsymbol{\beta}}^-) \geq \hat{w}((\tilde{\boldsymbol{\beta}}^-,\tilde{\boldsymbol{\beta}}^+) + \nabla\hat{w}(\tilde{\boldsymbol{\beta}}^{(t+1)+},\tilde{\boldsymbol{\beta}}^{(t+1)-})^\top\begin{pmatrix}\tilde{\boldsymbol{\beta}}^{t+}-\tilde{\boldsymbol{\beta}}^{(t+1)+}\\ \tilde{\boldsymbol{\beta}}^{t-}-\tilde{\boldsymbol{\beta}}^{(t+1)-}\end{pmatrix}. \tag{11}$$

Denote $h(\tilde{\boldsymbol{\beta}}) = w(\tilde{\boldsymbol{\beta}}) + \frac{\lambda}{2}r(\boldsymbol{\beta})$ and $h^*(\tilde{\boldsymbol{\beta}}) = w(\tilde{\boldsymbol{\beta}}) + \lambda r^*(\boldsymbol{\beta})$, where $r(\boldsymbol{\beta}) = \sum_{i=1}^m \frac{1}{\sqrt{|\beta_i^t|}}|\beta_i|$ and $r^*(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_{1/2}^{1/2} = \sum_{i=1}^m |\beta_i|^{1/2}$. Similarly, we define functions $\hat{r}(\cdot,\cdot)$ associated with $r(\cdot)$ and $\hat{r}^*(\cdot,\cdot)$ associated with $r^*(\cdot)$ as,

$$\hat{r}(\boldsymbol{\beta}^+,\boldsymbol{\beta}^-) = \sum_{i=1}^{m}\frac{1}{\sqrt{|\beta_i^t|}}(\beta_i^+ + \beta_i^-), \tag{12}$$

and

$$\hat{r}^*(\boldsymbol{\beta}^+,\boldsymbol{\beta}^-) = \sum_{i=0}^{m}(\beta_i^+ + \beta_i^-)^{1/2}. \tag{13}$$

Clearly, we have,

$$\hat{r}(\boldsymbol{\beta}^+,\boldsymbol{\beta}^-) = r(\boldsymbol{\beta}), \tag{14}$$

and

$$\hat{r}^*(\boldsymbol{\beta}^+,\boldsymbol{\beta}^-) = r^*(\boldsymbol{\beta}). \tag{15}$$

Further, we define $\hat{h}(\cdot,\cdot)$ associated with $h(\cdot)$ and $\hat{h}^*(\cdot,\cdot)$ associated with $h^*(\cdot)$ as,

$$\hat{h}(\tilde{\boldsymbol{\beta}}^+,\tilde{\boldsymbol{\beta}}^-) = \hat{w}(\tilde{\boldsymbol{\beta}}^+,\tilde{\boldsymbol{\beta}}^-) + \frac{\lambda}{2}\hat{r}(\boldsymbol{\beta}^+,\boldsymbol{\beta}^-), \tag{16}$$

$$\hat{h}^*(\tilde{\boldsymbol{\beta}}^+,\tilde{\boldsymbol{\beta}}^-) = \hat{w}(\tilde{\boldsymbol{\beta}}^+,\tilde{\boldsymbol{\beta}}^-) + \lambda\hat{r}^*(\boldsymbol{\beta}^+,\boldsymbol{\beta}^-). \tag{17}$$

Thus, it holds

$$\hat{h}(\tilde{\boldsymbol{\beta}}^+,\tilde{\boldsymbol{\beta}}^-) = h(\tilde{\boldsymbol{\beta}}), \tag{18}$$

and

$$\hat{h}^*(\tilde{\boldsymbol{\beta}}^+,\tilde{\boldsymbol{\beta}}^-) = h^*(\tilde{\boldsymbol{\beta}}). \tag{19}$$

Hence,

$$\tilde{\boldsymbol{\beta}}^{t+1} = \arg\min\ h(\tilde{\boldsymbol{\beta}}) = \arg\min\{w(\tilde{\boldsymbol{\beta}}) + \frac{\lambda}{2}r(\boldsymbol{\beta})\}. \tag{20}$$

$$\begin{aligned}(\tilde{\boldsymbol{\beta}}^{(t+1)+},\tilde{\boldsymbol{\beta}}^{(t+1)-}) &= \arg\min\ \hat{h}(\tilde{\boldsymbol{\beta}}^+,\tilde{\boldsymbol{\beta}}^-)\\ &= \arg\min\{\hat{w}(\tilde{\boldsymbol{\beta}}^+,\tilde{\boldsymbol{\beta}}^-) + \frac{\lambda}{2}\hat{r}(\boldsymbol{\beta}^+,\boldsymbol{\beta}^-)\}.\end{aligned} \tag{21}$$

From the optimality condition of $(\tilde{\boldsymbol{\beta}}^{(t+1)+},\tilde{\boldsymbol{\beta}}^{(t+1)-})$, we have

$$\begin{aligned}&\nabla\hat{h}(\tilde{\boldsymbol{\beta}}^+,\tilde{\boldsymbol{\beta}}^-)|_{(\tilde{\boldsymbol{\beta}}^{(t+1)+},\tilde{\boldsymbol{\beta}}^{(t+1)-})}\\ &= \nabla\hat{w}(\tilde{\boldsymbol{\beta}}^{(t+1)+},\tilde{\boldsymbol{\beta}}^{(t+1)-}) + \frac{\lambda}{2}\nabla\hat{r}(\boldsymbol{\beta}^{(t+1)+},\boldsymbol{\beta}^{(t+1)-}) = \mathbf{0}.\end{aligned} \tag{22}$$

Thus, combining Equations (11) and (22), we have

$$w(\tilde{\boldsymbol{\beta}}^t) \geq w(\tilde{\boldsymbol{\beta}}^{t+1}) + \frac{\lambda}{2}\nabla r(\boldsymbol{\beta}^{t+1})^\top(\tilde{\boldsymbol{\beta}}^{t+1} - \tilde{\boldsymbol{\beta}}^t), \tag{23}$$

and

$$\hat{w}(\tilde{\boldsymbol{\beta}}^{t+},\tilde{\boldsymbol{\beta}}^{t-}) \geq \hat{w}(\tilde{\boldsymbol{\beta}}^{(t+1)+},\tilde{\boldsymbol{\beta}}^{(t+1)-}) + \frac{\lambda}{2}\nabla\hat{r}(\boldsymbol{\beta}^{(t+1)+},\boldsymbol{\beta}^{(t+1)-})^\top\begin{pmatrix}\tilde{\boldsymbol{\beta}}^{(t+1)+} - \tilde{\boldsymbol{\beta}}^{t+}\\ \tilde{\boldsymbol{\beta}}^{(t+1)-} - \tilde{\boldsymbol{\beta}}^{t-}\end{pmatrix}. \tag{24}$$

In order to show the concavity of $\hat{r}^*$ with respect to $(\boldsymbol{\beta}^+,\boldsymbol{\beta}^-)$, we can easily compute the gradient based on its definition in Equation (13) as follows,

$$\nabla\hat{r}^*(\boldsymbol{\beta}^{t+},\boldsymbol{\beta}^{t-}) = \begin{pmatrix}\frac{1}{2}\cdot\frac{1}{\sqrt{\beta_1^{t+}+\beta_1^{t-}}}\\ \vdots\\ \frac{1}{2}\cdot\frac{1}{\sqrt{\beta_m^{t+}+\beta_m^{t-}}}\\ \frac{1}{2}\cdot\frac{1}{\sqrt{\beta_1^{t+}+\beta_1^{t-}}}\\ \vdots\\ \frac{1}{2}\cdot\frac{1}{\sqrt{\beta_m^{t+}+\beta_m^{t-}}}\end{pmatrix} \in \mathbf{R}^{2m}. \tag{25}$$

Since we know that,

$$\begin{aligned}\frac{\partial^2 \hat{r}^*(\boldsymbol{\beta}^+,\boldsymbol{\beta}^-)}{(\partial\beta_i^+)^2} &= \frac{\partial^2 \hat{r}^*(\boldsymbol{\beta}^+,\boldsymbol{\beta}^-)}{(\partial\beta_i^-)^2}\\ &= \frac{\partial^2 \hat{r}^*(\boldsymbol{\beta}^+,\boldsymbol{\beta}^-)}{\partial\beta_i^+\partial\beta_i^-}\\ &= -\frac{1}{4}(\beta_i^+ + \beta_i^-)^{-\frac{3}{2}}\\ &:= -\frac{a_i}{4},\end{aligned} \tag{26}$$

where, $a_i = (\beta_i^+ + \beta_i^-)^{-\frac{3}{2}} \geq 0, i = 1, \cdots, m$. Then, the Hessian of $\hat{r}^*$ is,

$$\nabla^2 \hat{r}^*(\boldsymbol{\beta}^+,\boldsymbol{\beta}^-) = \begin{pmatrix} -\frac{a_1}{4} & & & -\frac{a_1}{4} & & \\ & \ddots & & & \ddots & \\ & & -\frac{a_m}{4} & & & -\frac{a_m}{4} \\ -\frac{a_1}{4} & & & -\frac{a_1}{4} & & \\ & \ddots & & & \ddots & \\ & & -\frac{a_m}{4} & & & -\frac{a_m}{4} \end{pmatrix} \in \mathbf{R}^{2m\times 2m} \tag{27}$$

and $\forall \mathbf{u} = (u_1, \cdots, u_m)^\top, \mathbf{v} = (v_1, \cdots, v_m)^\top \in \mathbf{R}^m$,

$$(\mathbf{u}^\top, \mathbf{v}^\top)\nabla^2 \hat{r}^*(\boldsymbol{\beta}^+,\boldsymbol{\beta}^-)(\mathbf{u}^\top, \mathbf{v}^\top)^\top = -\frac{1}{4}\sum_{i=1}^{m} a_i(u_i + v_i)^2 \leq 0. \tag{28}$$

Therefore, $\hat{r}^*$ is concave with respect to $(\boldsymbol{\beta}^+, \boldsymbol{\beta}^-)$.
It follows directly from the concavity of $\hat{r}^*$ that

$$\hat{r}^*(\boldsymbol{\beta}^{(t+1)+}, \boldsymbol{\beta}^{(t+1)-}) \leq \hat{r}^*(\boldsymbol{\beta}^{t+}, \boldsymbol{\beta}^{t-}) + \nabla \hat{r}^*(\boldsymbol{\beta}^{t+}, \boldsymbol{\beta}^{t-})^\top \begin{pmatrix} \boldsymbol{\beta}^{(t+1)+} - \boldsymbol{\beta}^{t+} \\ \boldsymbol{\beta}^{(t+1)-} - \boldsymbol{\beta}^{t-} \end{pmatrix}, \tag{29}$$

which, in view of Equations (14) and (25), implies that

$$\begin{aligned} r^*(\boldsymbol{\beta}^{t+1}) &= \hat{r}^*(\boldsymbol{\beta}^{(t+1)+}, \boldsymbol{\beta}^{(t+1)-})\\ &\leq \hat{r}^*(\boldsymbol{\beta}^{t+}, \boldsymbol{\beta}^{t-}) + \sum_{i=1}^{m} \frac{(\beta_i^{(t+1)+} - \beta_i^{t+}) + (\beta_i^{(t+1)-} - \beta_i^{t-})}{2\sqrt{\beta_i^{t+} + \beta_i^{t-}}}\\ &= r^*(\boldsymbol{\beta}^t) + \frac{1}{2}\sum_{i=1}^{m} \frac{|\beta_i^{t+1}| - |\beta_i^t|}{\sqrt{\beta_i^{t+} + \beta_i^{t-}}}\\ &\leq r^*(\boldsymbol{\beta}^t) + \frac{1}{2}\sum_{i=1}^{m} \frac{|\beta_i^{t+1}| - \mathrm{sign}(\mathrm{fi}_\mathrm{i}^{\mathrm{t+1}})\mathrm{fi}_\mathrm{i}^\mathrm{t}}{\sqrt{\beta_i^{t+} + \beta_i^{t-}}}\\ &= r^*(\boldsymbol{\beta}^t) + \frac{1}{2}\begin{pmatrix} \frac{1}{\sqrt{|\beta_1^t|}}\mathrm{sign}(\mathrm{fi}_1^{\mathrm{t+1}}) \\ \vdots \\ \frac{1}{\sqrt{|\beta_m^t|}}\mathrm{sign}(\mathrm{fi}_\mathrm{m}^{\mathrm{t+1}}) \end{pmatrix}^\top \begin{pmatrix} \beta_1^{t+1} - \beta_1^t \\ \vdots \\ \beta_m^{t+1} - \beta_m^t \end{pmatrix}\\ &= r^*(\boldsymbol{\beta}^t) + \frac{1}{2}\nabla r(\boldsymbol{\beta}^{t+1})^\top(\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t). \end{aligned} \tag{30}$$

Multiplying λ on both sides of Equation (30) and subtracting from Equation (23), we have

$$\begin{aligned} h^*(\tilde{\boldsymbol{\beta}}^{t+1}) &= w(\tilde{\boldsymbol{\beta}}^{t+1}) + \lambda r^*(\boldsymbol{\beta}^{t+1}) \\ &\leq w(\tilde{\boldsymbol{\beta}}^{t}) + \lambda r^*(\boldsymbol{\beta}^{t}) \\ &= h^*(\tilde{\boldsymbol{\beta}}^{t}). \end{aligned} \tag{31}$$

From the definition of h^*, we can see $h^*(\tilde{\boldsymbol{\beta}}) \geq 0$, $\forall \tilde{\boldsymbol{\beta}} \in \mathbf{R}^{m+1}$. That is, h^* is monotonically decreasing function and bounded below. As stated in [22], by the LaSalleś Invariance Principle, $\{\tilde{\boldsymbol{\beta}}^t, t = 0, 1, 2, \cdots\}$ converges to the set of stationary points of $h^*(\tilde{\boldsymbol{\beta}})$ as $t \to \infty$. □

4. Performance Measure: Accuracy, Precision, and Recall

Since assessing the performance of a classifier is crucial in evaluating a classification model, we need to choose one or more proper performance measures.

For binary classification, a confusion matrix is usually used [24]. It summarizes the classification performance of a classifier in four categories: true positive (TP), false positive (FP), false negative (FN), and true negative (TN), as shown in Table 2. TP and TN outcomes are those classified correctly while FP and FN represent Type I error and Type II error, respectively.

Table 2. Confusion Matrix.

Actual	**Predicted**	
	Positive Class	**Negative Class**
Positive class	true positive (TP)	false negative (FN)
Negative class	false positive (FP)	true negative (TN)

A variety of common evaluation metrics can be derived from the confusion matrix, such as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{32}$$

and

$$\text{Error Rate} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \tag{33}$$

For imbalanced data, the application of accuracy and error rate results in a poor performance for the minority class, see [25].

Later, to cope with measure of classifiers for imbalance data, people develop some other evaluation metrics, to name a few, recall (also known as true positive rate (TPR), sensitivity), precision (also known as positive predictive value (PPV)), false positive rate (FPR), defined as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{34}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{35}$$

$$\text{False Positive Rate} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{36}$$

Thereafter, based on these metrics, the receiver operating characteristic (ROC) curve, the area under the ROC curve (AUROC, or just AUC), the Precision–Recall (PR) curve and the area under the PR curve (AUPRC) are developed. The ROC curve is a two-dimensional plot of classifier performance, which is obtained by plotting the TPR vs. the FPR for every possible classification threshold. It is useful for visualizing and evaluating the overall classification performance. To facilitate comparison, AUROC has been proposed, which summarizes the classification performance into a

single number. The PR curve is an alternative of the ROC curve that can visualize the performance of binary classification while AUPRC is its counterpart of AUROC.

As shown in Table 1, the data set is highly imbalanced. To balance between the risk of losing principal with potential investment opportunities, we care both the recall and precision. Therefore, AUPRC is more informative [25] in this case. Accuracy is also presented and we explain why it is not suitable here.

5. Experiments

We present the numerical results based on the historical loan information and data from LendingClub.

5.1. Data Description

LendingClub regularly updates the status of loans currently listed in data set available to download on a monthly basis and adds new loans data quarterly. In the data, the features include not only standard hard financial information commonly used by bank, such as annual income, debt-to-income ratio, FICO score range, but also non-standard information, such as description of the purpose of raising the loan, professional title. There are 151 features available in total. For more details of features available, we refere to the data dictionary provided by LendingClub (Data dictionary can be downloaded at https://resources.lendingclub.com/LCDataDictionary.xlsx). The number of features available may change over time.

The target variable of this experiment is loan status, while independent variables are carefully chosen from these 151 features. We take only the features can be described numerically into account, including numeric features and categorical features. Free text fields, such as emp_title, purpose, are removed. We finally take 62 features into consideration. To name a few,

- dti: Data to income ratio, a ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LendingClub loan, divided by the borrower's self-reported monthly income;
- emp_length: Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years;
- fico_range_high: The upper boundary range the borrower's FICO at loan origination belongs to;
- fico_range_low: The lower boundary range the borrower's FICO at loan origination belongs to;
- last_fico_range_high: The upper boundary range the borrower's last FICO pulled belongs to;
- last_fico_range_low: The lower boundary range the borrower's last FICO pulled belongs to;
- funded_amnt: The total amount committed to that loan at that point in time;
- last_pymnt_amnt: Last total payment amount received;
- max_bal_bc: Maximum current balance owed on all revolving accounts;
- inq_fi: Number of personal finance inquiries;
- zip_code: The first 3 numbers of the zip code provided by the borrower in the loan application;
- home_ownership: The home ownership status provided by the borrower during registration or obtained from the credit report. This is a categorical variable and possible values are: RENT, OWN, MORTGAGE, OTHER.

Here, we transform categorical features into binary features with dummy variables since they cannot be entered directly into a regression model and meaningfully interpreted. For more details about dummy variables, we refer to [26]. In addition, normalization of features is recommended to put different variables on the same scale in case there may be some features with far greater values than others, for instance, loan amount and annual income.

In this experiment, we choose data from the loans that already past the predetermined maturity. We consider loans with a 36-month maturity issued from 2013 to the first quarter of 2016 (2016Q1). The training sample size is 1000, while the testing sample size is 300. After gathering the data we first need to clean and prepare the data. Upon addressing missing data, special attention should be

paid since we may introduce bias at this step if the data are not missing at random. We transform date information to time length from the date to the day we perform this experiment. In particularly, the feature emp_length seems numeric, since it ranges from 0 to 10. However, since 0 means less than one year and 10 means ten or more years, it is actually a categorical feature. We transform such categorical features into binary features with dummy variables by replace a feature of c categories with $c-1$ dummy variables. Then, we apply normalization. Later, highly correlated predictors should be removed in order to reduce multicollinearity. Finally, we split the data into training sample set and testing sample set for in-sample tests and out-of-sample tests, separately.

As mentioned above, the datasets we considerate are highly imbalanced. Table 3 shows the imbalance ratios of sample sets, defined as the ratio of the number of instances in major class to the number of examples in the minority class. Here, the major class is Fully Paid; the minority class is Not Fully Paid.

Table 3. Imbalance ratio.

Dataset	Imbalance Ratio			
	2013	**2014**	**2015**	**2016Q1**
Train	5.250005	3.854367	2.521127	
In-sample test	5.666653	3.109588	2.797469	
Out-of-sample test		3.285712	2.191488	3.285712

5.2. Numerical Results

This section contains training, in-sample test, and out-of-sample test results for the year 2013, 2014, and 2015. We performed in-sample tests with instances sampled from the training sample set, while we conducted out-of-sample tests with examples sampled from the next period.

Here, we chose five different values for the penalty factor on the negative class, $\theta = 1, 2, 3, 4, 5$, based on the imbalance ratio of the dataset and five different values for the regularization parameter, $\lambda = 0, 10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}$, based on the value of loss function and the regularization term. When $\theta = 1$ and $\lambda = 0$, the model reduces to a standard logistic regression.

Figures 1–3 show the AUPRC, accuracy, precision, and recall results of training, in-sample test, and out-of-sample test for 2013, 2014, and 2015. We also present the AUPRC results in Table 4.

From these scatter plots, we can see accuracy performs poorly for imbalanced data. Tests with nearly the same accuracy may be far different in the number of FP samples and that of FN samples. Accuracy only shows the percentage of samples correctly classified and do not distinguish between FP and FN samples, which makes it simply does not work in our case.

As mentioned above, the probability of making a Type I error and that of making a Type II error cannot be reduced simultaneously for a given sample. Recall and precision in general change in opposite directions. As shown in the figures, for a fixed λ, precision results tend to increase with the increase of θ at the cost of the reduction in recall. Investors that are more risk-averse could apply a higher θ to keep the principal safer, while it may cause loss of investment opportunities.

Since the number of features taken into consideration is considerable, overfitting may happen under the standard logistic regression. Regularization could help to reduce the chance of, or the amount of, overfitting. As shown in Table 4, we present the AUPRC results of training, in-sample test, and out-of-sample test for 2013, 2014, and 2015. For a fixed θ, a higher regularization parameter λ in general yields higher out-of-sample AUPRC.

Table 4. AUPRC results of training, in-sample test, and out-of-sample test for 2013, 2014, and 2015.

θ	λ	2013			2014			2015		
		Training	**In-Sample Test**	**Out-of-Sample Test**	**Training**	**In-Sample Test**	**Out-of-Sample Test**	**Training**	**In-Sample Test**	**Out-of-Sample Test**
1.0	0	0.9952	0.8783	0.8170	0.9889	0.7938	0.7440	0.9771	0.7752	0.8074
	1.0×10^{-10}	0.9952	0.8783	0.8197	0.9888	0.7947	0.7445	0.9776	0.7745	0.8068
	1.0×10^{-8}	0.9952	0.8798	0.8179	0.9889	0.7977	0.7422	0.9773	0.7731	0.8127
	1.0×10^{-6}	0.9951	0.8836	0.8221	0.9890	0.7993	0.7446	0.9773	0.7735	0.8183
	1.0×10^{-4}	0.9936	0.9048	0.8332	0.9875	0.7969	0.7684	0.9748	0.8058	0.8478
2.0	0	0.9956	0.8792	0.8210	0.9895	0.7933	0.7446	0.9790	0.7744	0.8094
	1.0×10^{-10}	0.9957	0.8791	0.8171	0.9896	0.7935	0.7450	0.9789	0.7755	0.8095
	1.0×10^{-8}	0.9956	0.8801	0.8205	0.9895	0.7963	0.7428	0.9789	0.7736	0.8137
	1.0×10^{-6}	0.9957	0.8853	0.8232	0.9895	0.7981	0.7408	0.9789	0.7749	0.8204
	1.0×10^{-4}	0.9955	0.9019	0.8345	0.9893	0.8031	0.7640	0.9786	0.7989	0.8438
3.0	0	0.9958	0.8793	0.8237	0.9898	0.7881	0.7453	0.9793	0.7730	0.8125
	1.0×10^{-10}	0.9958	0.8790	0.8181	0.9896	0.7926	0.7442	0.9792	0.7745	0.8113
	1.0×10^{-8}	0.9958	0.8799	0.8235	0.9898	0.7962	0.7413	0.9792	0.7705	0.8153
	1.0×10^{-6}	0.9959	0.8861	0.8230	0.9897	0.7965	0.7437	0.9792	0.7780	0.8203
	1.0×10^{-4}	0.9956	0.8975	0.8370	0.9897	0.8028	0.7627	0.9791	0.7954	0.8466
4.0	0	0.9959	0.8786	0.8213	0.9898	0.7928	0.7422	0.9794	0.7707	0.8112
	1.0×10^{-10}	0.9959	0.8790	0.8239	0.9898	0.7882	0.7454	0.9793	0.7717	0.8117
	1.0×10^{-8}	0.9959	0.8805	0.8218	0.9898	0.7957	0.7431	0.9795	0.7736	0.8162
	1.0×10^{-6}	0.9958	0.8876	0.8268	0.9898	0.7965	0.7420	0.9794	0.7814	0.8211
	1.0×10^{-4}	0.9956	0.8998	0.8348	0.9897	0.7994	0.7613	0.9790	0.7936	0.8464
5.0	0	0.9959	0.8791	0.8236	0.9898	0.7937	0.7454	0.9794	0.7710	0.8120
	1.0×10^{-10}	0.9959	0.8791	0.8226	0.9898	0.7891	0.7423	0.9795	0.7750	0.8124
	1.0×10^{-8}	0.9960	0.8803	0.8218	0.9898	0.7961	0.7421	0.9793	0.7705	0.8160
	1.0×10^{-6}	0.9959	0.8874	0.8252	0.9898	0.7965	0.7421	0.9794	0.7825	0.8214
	1.0×10^{-4}	0.9956	0.8983	0.8341	0.9895	0.7976	0.7607	0.9790	0.7934	0.8477

The first column, θ, is the penalty factor on the negative class. The second column, λ, is the regularization parameter. Column 3–5, 6–8, 9–11 show the AUPRC results of training, in-sample test, and out-of-sample test, for 2013, 2014, 2015, respectively.

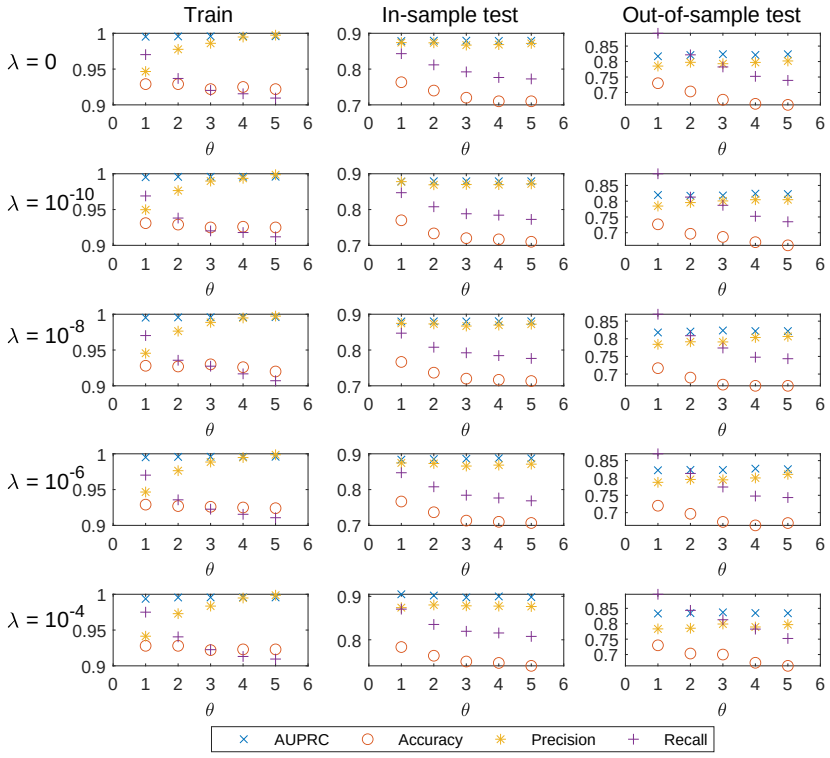

Figure 1. Area under the Precision–Recall curve (AUPRC), accuracy, precision, and recall results of training, in-sample test, and out-of-sample test for 2013.

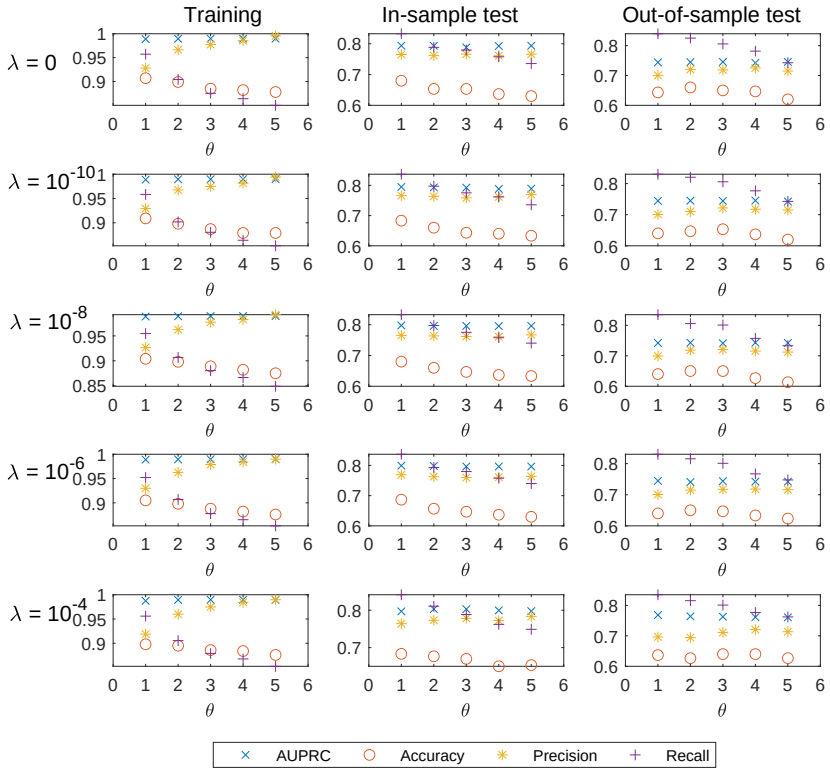

Figure 2. AUPRC, accuracy, precision, and recall results of training, in-sample test, and out-of-sample test for 2014.

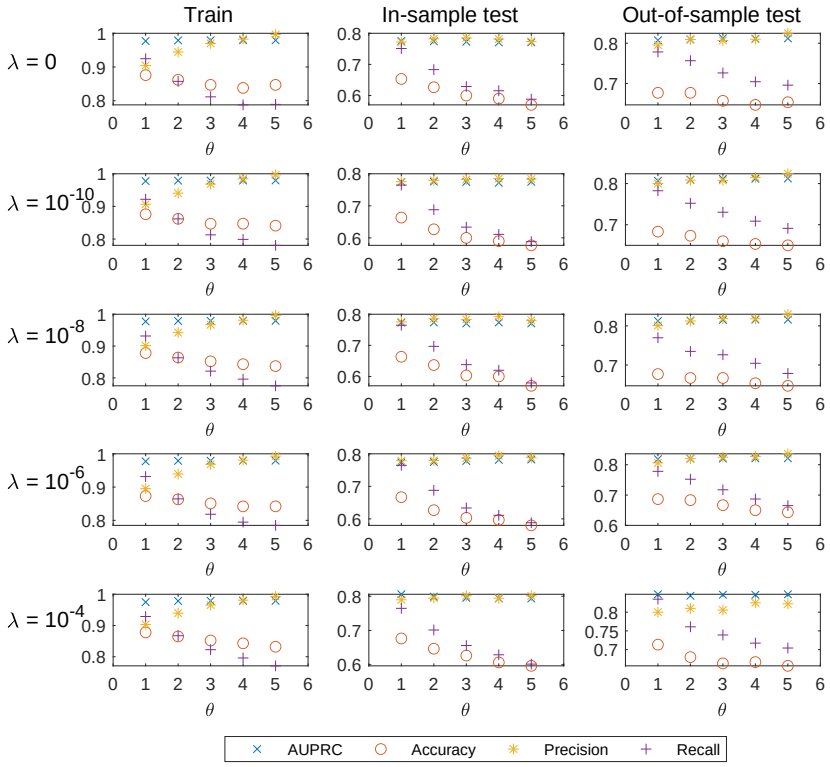

Figure 3. AUPRC, accuracy, precision, and recall results of training, in-sample test, and out-of-sample test for 2015.

6. Discussion

The objective of this paper was to provide a method for investors in the P2PL marketplace to perform default prediction, where there exists a high-level of information asymmetry. We considered LendingClub since the availability of historical data. Since investors in P2PL are mostly individuals and small businesses. When involved in P2PL marketplace, investors are frequently adversely affected by the information asymmetry. Additionally, not every investor has a solid background in investment or quantitative finance. This makes a relatively easy and straightforward model needed.

We propose an $L_{1/2}$-regularized weighted logistic model. Via only adjusting the penalty factor θ and the regularization parameter λ, investors can find a trade-off between the risk of losing principal and that of losing potential investment opportunities according to their own risk preferences and lessen the chance of, or amount of, overfitting in the meantime.

Numerical experiment shows that a higher regularization parameter yields better out-of-sample AUPRC and investors that are more risk-averse could lower the risk of losing principal at the cost of potential investment opportunities by increasing the penalty factor on the negative class according to their own risk preferences. This default prediction could help investors protect their profits and principle in the disadvantage of information asymmetry.

7. Limitations and Further Research

Since we solve the proposed model with an iterative algorithm, it has the shortcomings of longer calculation, especially when the sample size is large. Further, high performance computing could be applied to improve computing efficiency.

Author Contributions: Methodology, X.W., data curation, X.W., data curation, X.W., writing–original draft, X.W., writing–review and editing, Y.L., visualization, X.W., supervision, B.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (11971092).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lynn, T.; Mooney, J.G.; Rosati, P.; Cummins, M. *Disrupting Finance: Fintech and Strategy in the 21st Century*; Springer: Cham, Switzerland, 2018.
2. Lai, L.S.; Turban, E. Groups formation and operations in the Web 2.0 environment and social networks. *Group Decis. Negot.* **2008**, *17*, 387–402. [CrossRef]
3. Smith, A.M. SEC Cease-and-Desist Orders. *Adm. Law Rev.* **1999**, *51*, 1197–1228.
4. Yang, H. Comprehensive Evaluation of Online Peer-to-Peer Lending on the Province-Level Regions in China Based on Generalized Principle Component Analysis. *Open J. Bus. Manag.* **2016**, *4*, 171–176. [CrossRef]
5. Wang, H.; Greiner, M.; Aronson, J.E. People-to-people lending: The emerging e-commerce transformation of a financial market. In *Value Creation in E-business Management*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 182–195.
6. Barasinska, N.; Schäfer, D. Is crowdfunding different? Evidence on the relation between gender and funding success from a German peer-to-peer lending platform. *Ger. Econ. Rev.* **2014**, *15*, 436–452. [CrossRef]
7. Xia, Y. A Novel Reject Inference Model Using Outlier Detection and Gradient Boosting Technique in Peer-to-Peer Lending. *IEEE Access* **2019**, *7*, 92893–92907. [CrossRef]
8. Byanjankar, A.; Heikkilä, M.; Mezei, J. Predicting credit risk in peer-to-peer lending: A neural network approach. In Proceedings of the 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, South Africa, 7–10 December 2015; pp. 719–725.
9. Jiang, C.; Wang, Z.; Wang, R.; Ding, Y. Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Ann. Oper. Res.* **2018**, *266*, 511–529. [CrossRef]
10. Kim, A.; Cho, S.B. An ensemble semi-supervised learning method for predicting defaults in social lending. *Eng. Appl. Artif. Intell.* **2019**, *81*, 193–199. [CrossRef]
11. Wei, X.; Gotoh, J.Y.; Uryasev, S. Peer-To-Peer Lending: Classification in the Loan Application Process. *Risks* **2018**, *6*, 129. [CrossRef]
12. Wei, Z.; Lin, M. Market mechanisms in online peer-to-peer lending. *Manag. Sci.* **2016**, *63*, 4236–4257. [CrossRef]
13. Liu, H.L.; Chen, H.Z.; Ding, Y.J.; Zhang, X. Research on Investment Model of Internet Financial Loan Platform. In Proceedings of the International Conference on Artificial Intelligence and Computing Science, Hangzhou, China, 24–25 May 2019; DEStech Publications: Lancaster, PA, USA; pp. 311–315
14. Cho, P.; Chang, W.; Song, J.W. Application of instance-based entropy fuzzy support vector machine in peer-to-peer lending investment decision. *IEEE Access* **2019**, *7*, 16925–16939. [CrossRef]
15. Ren, K.; Malik, A. Investment Recommendation System for Low-Liquidity Online Peer to Peer Lending (P2PL) Marketplaces. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Melbourne, Australia, 11–15 February 2019; ACM: New York, NY, USA, 2019; pp. 510–518.
16. Calabrese, R.; Osmetti, S.A. Modelling small and medium enterprise loan defaults as rare events: the generalized extreme value regression model. *J. Appl. Stat.* **2013**, *40*, 1172–1188. [CrossRef]
17. Ma, L.; Zhao, X.; Zhou, Z.; Liu, Y. A new aspect on P2P online lending default prediction using meta-level phone usage data in China. *Decis. Support Syst.* **2018**, *111*, 60–71. [CrossRef]
18. Lin, M.; Prabhala, N.R.; Viswanathan, S. Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Manag. Sci.* **2013**, *59*, 17–35. [CrossRef]
19. Ge, R.; Feng, J.; Gu, B.; Zhang, P. Predicting and deterring default with social media information in peer-to-peer lending. *J. Manag. Inf. Syst.* **2017**, *34*, 401–424. [CrossRef]
20. Smith, C.E. If it's not broken, don't fix it: The SEC's regulation of peer-to-peer lending. *Bus. Law Brief* **2009**, *6*, 21.
21. Tsai, K.; Ramiah, S.; Singh, S. Peer Lending Risk Predictor. CS229 Autumn. 2014. Available online: https://www.researchgate.net/profile/Sudhanshu_Singh8/publication/269699712_Peer_Lending_Risk_Predictor/links/549321420cf286fe3125b7d3/Peer-Lending-Risk-Predictor.pdf (accessed on 10 December 2016).
22. Xu, Z.B.; Zhang, H.; Wang, Y.; Chang, X.Y.; Liang, Y. $L_{1/2}$ regularization. *Sci. China Inf. Sci.* **2010**, *53*, 1159–1169. [CrossRef]

23. Zeng, J.; Lin, S.; Wang, Y.; Xu, Z. $L_{1/2}$ regularization: Convergence of iterative half thresholding algorithm. *IEEE Trans. Signal Process.* **2014**, *62*, 2317–2329. [CrossRef]
24. Ting, K. Confusion Matrix. *Encycl. Mach. Learn.* **2010**, *1*, 209.
25. Fayzrakhmanov, R.; Kulikov, A.; Repp, P. The Difference Between Precision-recall and ROC Curves for Evaluating the Performance of Credit Card Fraud Detection Models. In Proceedings of the 6th International Conference on Applied Innovations in IT, Koethen, Germany, 13 May 2018; pp. 17–22.
26. Suits, D.B. Use of dummy variables in regression equations. *J. Am. Stat. Assoc.* **1957**, *52*, 548–551. [CrossRef]