Expert Systems with Applications 42 (2015) 4621–4631

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Risk assessment in social lending via random forests

Milad Malekipirbazari, Vural Aksakalli *

Istanbul Sehir University, Istanbul 34662, Turkey

ARTICLE INFO

Article history:
Available online 14 February 2015

Keywords:
Peer-to-peer lending
Social lending
Risk assessment
Machine learning
Random forest

ABSTRACT

With the advance of electronic commerce and social platforms, social lending (also known as peer-to-peer lending) has emerged as a viable platform where lenders and borrowers can do business without the help of institutional intermediaries such as banks. Social lending has gained significant momentum recently, with some platforms reaching multi-billion dollar loan circulation in a short amount of time. On the other hand, sustainability and possible widespread adoption of such platforms depend heavily on reliable risk attribution to individual borrowers. For this purpose, we propose a random forest (RF) based classification method for predicting borrower status. Our results on data from the popular social lending platform Lending Club (LC) indicate the RF-based method outperforms the FICO credit scores as well as LC grades in identification of good borrowers.

1. Introduction

Social lending, also known as peer-to-peer (P2P) lending, is emerging as an alternative to banks where individual members lend and borrow money using an online trading platform without the help of official financial intermediaries such as banks. The attractive feature of doing business on a peered platform is the higher potential of mutual profitability. Borrowers can obtain loans at lower interest rates and lenders can loan money at better rates than what they can get from a bank. In particular, via social lending, lenders can find a multitude of potential borrowers and choose among them the ones they wish to lend. Since the ultimate savers are predominantly consumers, and consumers are the individuals who are actually lending in the social lending model, there is no need to increase the liquidity of the loans by securitizing them. Since social lending is powered by the Internet, it would not take much effort to connect small communities such as towns, religious, or ethnic groups for the purpose of intra-community lending and borrowing.

The popular social lending platforms currently in use today are the U.S.-based Prosper[1] and Lending Club Corp.,[2] UK-based Zopa Ltd.,[3] and Germany-based Smava GmbH.[4] All of these social platforms rely on the credit scores provided by a cooperating credit reporting agency; Experian, TransUnion LLC, Equifax Inc., and Schufa Holding AG respectively. The popularity of these platforms is growing as recently indicated by Lending Club (LC) which has reached 6.2 billion USD in total loans by January 2015 and transformed into a 8.5 billion USD publicly-traded company, becoming the world's largest social lending platform (LendingClub.com, 2015).

Our standpoint in this work, which is consistent with other studies in this context, is that even though social lenders can base their investment strategy on the traditional financial credit scores provided by external agencies, available data suggest that social lending tends to have different dynamics when compared to traditional lending. For instance, the distribution of lenders' bids on social loan listings when indexed by time follows a power law (Rodgers & Zheng, 2002), an indication of a herding behavior. Assume that lenders and loan listings are denoted by nodes and an edge between them denotes that the lender is interested in the corresponding listing. Since the distribution of bids indicates a bias towards highly connected nodes, this in reality means that once a loan listing has a hundred or more lenders bid on it, then that specific listing is more likely to attract more and more lenders. This in turn makes the corresponding listing more likely to get funded in the end due to high lender interest.

A comprehensive analysis of LC loan data by Emekter, Tu, Jirasakuldech, and Lud (2015) reveals two key findings:

1. There exists a selection bias in the sense that high-income borrowers with the highest FICO credit scores[5] do not borrow from LC. In particular, top one third of the consumers with respect to FICO scores do not create any loan listings on LC.

* Corresponding author.
E-mail addresses: miladmalekipirbazari@std.sehir.edu.tr (M. Malekipirbazari), aksakalli@sehir.edu.tr (V. Aksakalli).

[1] http://www.prosper.com.
[2] http://www.lendingclub.com.
[3] http://www.zopa.com.
[4] http://www.smava.de.

[5] FICO, a publicly traded corporation, produces scoring models that are most commonly used and distributed by TransUnion, Equifax, and Experian.

http://dx.doi.org/10.1016/j.eswa.2015.02.001
0957-4174/

2. Higher interest rates charged on the higher risk borrowers are not worth the risk. Specifically, higher rates charged for the borrowers with low LC's own credit grade are not high enough to overcome the greater default risk that the lenders take.

The above two findings imply that, from a profitability point of view, identifying the ''good borrowers'', i.e., those who will pay back their loan in full within due time, is of great importance for investors participating in social lending. Profitability of social investors, on the other hand, is a critical component in continued interest in social lending as well as overall sustainability of the social lending market. In this regard, subsequent to a risk and return efficiency analysis, Emekter et al. (2015) suggest that ''the lenders would be better off to lend only to the safest borrowers with the highest LC grades''. Despite this suggestion, we show in this work that even borrowers with the highest FICO scores or LC grades are not necessarily good borrowers, which in turn indicate that traditional financial score metrics are not well-equipped to capture the non-conventional dynamics prevalent in social lending.

In order to improve identification of good borrowers within the context of social lending, this study proposes and presents comparisons of different machine learning methods including random forests (RFs), support vector machines (SVM), logistic regression (LR), and k-nearest neighbor (k-NN) classifiers. Our computational results on LC data between January 2012 and September 2014 for a total loan amount of about one billion USD indicate that random forests outperform the other classification methods and stand as a scalable and powerful approach for predicting borrower status. In fact, an empirical comparison reveals that RFs significantly outperform both FICO scores and LC grades in identification of the best borrowers in terms of low default probability.

The rest of this manuscript is organized as follows: Section 2 provides a brief literature review on social lending. In Section 3, we introduce basics of social lending, describe the financial features used in prediction, and provide an exploratory data analysis. Section 4 presents the classifiers and Section 5 provides a comprehensive experimental comparison. Our summary and conclusions are presented in Section 6.

2. Related work

An application of machine learning principles in social lending is the use of Gaussian mixture models on the Prosper data set containing loan transactions between November 2005 and December 2008 (Lopez, 2009). An interesting finding therein is that if an individual with a high-risk FICO score belongs to a trusted social community, then this individual's social membership can still help secure a loan. Thus, even though a high-risk credit score usually means lack of access to traditional bank-mediated financial markets, a positive social feature can outweigh a highly negative financial feature in socially mediated markets.

Complex behavioral dynamics further complicate the social lending process. For example, the simple auction mechanism used in some social lending platforms can lead to unpredictable payments for the borrower. An incentive compatible mechanism might be more suitable to eliminate this inefficiency where lenders report their true interest rate and do not change their rate dynamically (Chen, Ghosh, & Lambert, 2009). Otherwise, such inefficiencies enable users with adversarial interests to use the lending platform as an arbitrage opportunity: borrow at 10% and then loan at 20% (Steelman, 2006).

The notion of groups was introduced into social lending with Prosper. Users of this platform can form groups around an affinity that all members share such as a certain topical interest, geographic location, a peer group, or simply around the reasons to borrow. Groups have leaders that act as mediators of loan activity. This mediation can be in the form of pre-evaluating group members, endorsing potential borrowers, inverting and diffusing the risk of a particular group member default among all group members, or encouraging all members to proactively screen new members and apply peer pressure. Empirical studies show that when a group leader in a lending platform mediates the group actively, the risk factor drops considerably. In addition, if a group leader recommends a loan listing put together by one of the group members, this endorsement increases the chance of the loan being issued and also decreases the final interest rate (Berger & Gleisner, 2009).

There exist several studies proposing a set of guidelines in order to make purely rational investment decisions in social lending. In one such study on Prosper loan data that includes loan transactions between November 2005 and March 2007, irrespective of the financial credit rating categories,[6] three simple rules help decrease the risk of a default (Klafft, 2008). These investment rules are as follows:

1. Invest only in borrowers without any delinquent accounts.
2. Invest only in borrowers that satisfy Rule 1 and that have a debt-to-income (DTI) ratio less than 20%.
3. Invest in borrowers that satisfy Rule 2 and that have no credit inquiry reports during the last 6 months.

In studies conducted on social communities, herding (denser clustering following a power law regime) effects usually prevail (Gao & Feng, 2014; Lee & Lee, 2012; Yum, Lee, & Chae, 2012). Empirical studies show that the tendency of an individual to join a given community is effected by the number of friends in this community and the inter-connectedness of this individual's friends within the community. Such behavioral bias also exist in investment decisions of lenders at Prosper. The loan data between 2006 and 2008 show that previous lender decisions effected subsequent lender decisions and lender decisions were not made purely rationally (Shen, Krumme, & Lippman, 2010). For the interested reader, there exist other real-world networks (such as airports and power grid transmission lines) and other social networks (such as DBLP and LiveJournal) that also exhibit a herding behavior (Amaral, Scala, Barthelemy, & Stanley, 2000; Backstrom, Huttenlocher, Kleinberg, & Lan, 2006).

The closest study to ours is the work of Emekter et al. (2015) where the authors analyze LC data between May 2007 and June 2012 and present a logistic regression (LR) model for predicting default probability of a borrower. Their model includes FICO scores as well as LC grades in default prediction. In contrast, our study uses all the available financial features other than the FICO scores and LC grades in order to assess the relevance and prediction power of these two metrics in social lending. Nonetheless, we show in Section 4 that one can get much better prediction accuracy using random forests compared to LR even with the exact same features used in building this LR model.

3. Overview and data analysis

3.1. Social lending overview

The LC social lending platform works as follows:

[6] Prosper grades its individual platform users into credit grade buckets in the increasing risk order as AA, A, B, C, D, E, and HR (high-risk) depending purely on credit scores assigned by Experian.