DOI: 10.16538/j.cnki.jsufe.2019.02.007

行为轨迹及社交信息能预测用户是否再借款吗？
——基于网络借贷的实证分析

黄　静，缪世磊
（上海师范大学　商学院，上海　200234）

摘　要：网络借贷具有高频次、反复借贷的特点，用户往往具有再次借款的需求，利用网络平台累积的各种信息，特别是用户的“行为轨迹”和“社会交往”数据，对平台上具有“再次借款”可能的优质用户进行挖掘，不仅能提高平台的运营效率，也能促使网络借贷市场平稳长远的发展。文章采用XGBoost算法，利用网络借贷平台上2.6万个用户68万多条数据，首创性地建立了用户“是否再次借款”的预测模型，并对用户关键特征进行可视化分析。主要结论为：网络借贷中预测用户是否申请再借款，用户提供的“硬信息”（个人基本信息）已不具有信号揭示作用，用户在平台上的借贷和消费的“行为轨迹”信息以及“社会交往”信息，更具有信任信号的揭示作用。如果在平台上积极维护个人信息、保持良好的还款记录、维持良好网络社交的用户，再次申请借款的可能性就很高。

关键词：网络借贷；行为轨迹；社会交往；信号揭示
中图分类号：F830.5　　**文献标识码：**A　　**文章编号：**1009-0150(2019)02-0093-13

一、引　言

网络借贷给借款人和投资人提供了直接进行资金对接的平台，不仅为资金需求者（个人或小微企业借款人）开启了新的融资渠道，也为资金所有者（投资人）带来较为稳定的收益，在提高借贷对接效率的同时，也降低了交易成本。自2006年成立了第一家“宜信”公司以来，我国网络借贷的平台数量迅猛增长，交易规模也迅速扩大。根据行业报告的数据显示[①]，截至2018年3月底，历史累计成交量为68 027.09亿元，3月份行业成交量为1 915.65亿元，正常运营平台数为1 883家。虽然我国网络借贷发展速度迅猛，但是目前我国网络借贷市场仍然处于粗放式发展阶段，许多借贷平台的运营效率不高，很多投资也无法准确地找到合适的借款对象，借贷成功率较低。

与传统线下借贷相比，网络借贷具有小额化和高频次的特点。据融360统计，在消费金融领域，70%以上的用户是二次或多次借款的用户。也就意味着，用户在网络借贷平台成功借款后，往往会进行第二次或多次借款（本文统称为“再借款”）。具有再借款意图的用户，通常会更在意自身的信用和声誉，向平台或投资人揭示其信任信号，其道德风险会更低，往往成为网络借贷平台的优质客户，可为平台创造更多的价值。因此，对网络借贷平台的用户是否会“再借款”

收稿日期：2018-12-01
基金项目：上海市哲学社会科学规划课题“基于‘房住不炒’视角的房地产市场参与主体行为分析及长效机制研究（2018BJB024）”。
作者简介：黄　静（1977－），女，江西南城人，上海师范大学商学院副教授、博士；
缪世磊（1992－），男，安徽合肥人，上海师范大学商学院硕士研究生。

①数据来源于融希财网出具的网贷行业月度报告，http://www.csai.cn/p2pzixun/1264720.html，2018年4月20日。

进行建模预测及特征识别，挖掘这部分优质客户在平台上的信息，以提取高价值客户特征，无疑将提高借贷成功率，从而提高网络借贷平台的运营效率。

网络借贷具有借贷双方互不相识、通过互联网直接交易以及资金交易无抵押品等特点，如何解决信息不对称问题以保持平台平稳、健康及高效运行，是实务界和理论界关注的焦点问题。投资人和借款人信息不对称，借款人具有信息优势可能产生道德风险，投资人处于信息劣势可能产生逆向选择（Yum等，2012；廖理等，2014）。因此，网络借贷平台如何加强对借款人的信息甄别，如何进行审核和筛选成为关键。目前，各平台对借款人的审核方式主要还是采用传统的认证指标，例如，通过收入认证指标、房产和车产认证指标、工作认证指标等方式（王会娟和廖理，2014）。毋庸置疑，传统的认证机制能够在一定程度上缓解由于信息不对称导致的道德风险问题。

然而，互联网金融与传统金融相比，其优势在于互联网渠道的网络借贷平台的运作过程中自动积累了大量的信息。除借贷双方主动提供的个人信息之外，借贷平台上记录了所有的借贷消费信息，以及大量的借贷者的“行动轨迹”大数据，有些网络借贷平台给借贷双方提供了“社会交往”功能，记录了用户的社会网络关系。如何利用这些大数据，对再借款客户信息进行挖掘分析，以弱化网络借贷中的信息不对称问题，提高平台的运营效率，是值得我们研究的重点问题。国内现有研究侧重于网络借贷的模式介绍、影响因素、风险及其监管问题的研究，没有区分首次借款用户和再次借款用户，忽略了对网络借贷平台上自我记录所形成的大数据进行挖掘研究，特别是再借款人的“行动轨迹”以及平台社交产生的“社会资本”信息等大数据。

本文首先采用融360网络借贷平台提供的千万量级的用户行动轨迹和社交信息大数据，利用人工智能中鲁棒性非常好并且可解释性较强的XGBoost算法，建立“是否申请再借款”的评分模型；然后，基于模型结果，对“是否申请再借款”的影响特征进行重要性排序，依据模型打分，提取重要的影响因素，并用可视化方法展示结果，以回答“用户行为轨迹及社交信息能预测是否再借款吗？”这一问题。本文首创性地建立“是否申请再借款”评分模型，利用网络借贷平台中用户的“行为轨迹”数据和“社会资本”数据，挖掘网络借贷平台用户再借款的信号揭示行为特征，可以帮助网络借贷平台筛选优质客户，以提高平台的运营效率及借贷成功率，这将有利于网络借贷市场的长远及稳定发展。

二、文献综述

既有的网络借贷研究，按借贷双方、借贷平台、监管机制等参与主体可以分为三大类：第一类是关于网络借贷行为的影响因素研究，包括借款人借款成功的决定因素以及借款利率高低的影响因素分析（Klafft，2008；Herzenstein等，2008；Iyer等，2009），也包括贷款人的风险识别能力的分析（廖理等，2014；蒋翠清等，2017）；第二类是网络借贷平台的监管问题研究（Davis和Gelpern，2010；Slattery，2013；张国文，2014；杨振能，2014）；第三类则集中于介绍网络借贷平台的运营模式以及存在的问题（王硕，2015；莫易娴，2011；王修华等，2016）。

网络借贷没有抵押担保，存在严重的信息不对称问题，平台或投资者完全依靠信息来判断贷款人的信用水平以规避道德风险。既有研究中，把网络借贷中借款人揭示其信任的信息归为两类，分别被称为“硬信息”和“软信息”。

所谓的“硬信息”是指借款人必须按照平台要求提供的可确认、可验证的信息，主要包括基本信息、资产信息和工作信息。例如，基本信息包括身份证号、性别、年龄、学历以及婚姻状况

等；资产信息包含收入、车产、房产以及贷款等资产和负债情况；工作信息一般包含借款人所从事的行业、公司规模、工作年限以及工作城市等（王会娟和何琳，2014）。“硬信息”是平台评估借款人信用等级的主要依据，成为网络借款是否成功的关键因素（Herzenstein等，2008；Weiss等，2010；Iyer等，2009）。王会娟和廖理（2014）利用“人人贷”平台的数据进行分析，结果发现平台信用认证机制认定的信用级别越高，借款成功概率就越高，并且借款成本也越低。平台在进行信用级别认证时，收入情况、工作情况、车产和房产等认证指标对信用认证级别影响很大。Freedman和Jin（2008）研究认为借款人在平台上提供的财务方面的信息越多，则越容易获取贷款，贷款人的年龄（Pope和Sydnor，2011）、性别、种族（Herzenstein等，2008）等特征都会影响借贷行为。Pope和Sydnor（2011）研究认为，35岁以下的人要比35–60岁的群体借款成功的可能性高出0.4–0.9个百分点，60岁及以上的人借款成功的可能性要比35–60岁群体低1.1–2.3个百分点。孙武军和樊小莹（2016）利用“人人贷”数据研究了借款人的从业经历与教育背景对网络借贷成功率的影响，结果认为借款人的学历越高或工作经历越丰富，其借贷成功率越高。

所谓的“软信息”是指借款人自愿披露的关于自己个人的品质、经历、性格、借款原因以及未来希望等方面的描述性信息，通过叙述性的语言由借款人自愿性地表述出来（李焰等，2014）。学者们的研究表明，描述性的软信息在一定程度上预示着借款人的还款行为，促使投资者建立对其的信任，有助于提高贷款成功率（Herzenstein等，2011；李焰等，2014）。网络借贷中存在严重的信息不对称，借款人通过“硬信息”和“软信息”综合起来揭示其可信任信号，可以提高借贷市场的信息透明度，投资人通过这些信息以判断借款人的整体信用情况，有助于抑制信息扭曲，在一定程度上可以解决信息不对称问题（Larrimore等，2011）。

其实，网络借贷中借助于互联网平台的优势，除积累了用户主动提供的大量的“硬信息”和“软信息”之外，还记录了用户在平台上的“行为轨迹”数据和“社会交往”数据。许多网络借贷平台给用户之间提供了相互关注和交互活动的功能，产生了大量的社会关系交互数据，可以作为社会资本分析的数据。Greiner和Wang（2009）研究了借贷交易行为与社会资本之间的关系，研究发现用户拥有越多的社会资本，其借款的成功率也就越大，借款利率也更低。可以看出，借款人在网络借贷平台上的历史信息、社会资本及其行为特征，共同形成借款人的声誉，这类似于网络借贷平台的无形资产，对此信息的挖掘可降低借贷双方的谈判成本（Lin等，2013），也可以降低平台对借款人的监督成本（王博等，2017）。

综上所述，现有研究利用借款人提供的“硬信息”或“软信息”、用户平台上积累的“行为轨迹”和“社会交往”数据，重点研究和识别了网络借贷中的借贷成功因素、利率高低的决定因素以及违约的影响因素，形成了丰富的研究成果。然而，现有文献相对忽视了网络借贷中“频次高”和“重复借贷”的特点，没有区分首次借款和再次（或多次）借款者的行为差异，导致现有研究中缺乏针对“再借款”这一类优质和忠实客户进行信息挖掘的相关研究。本文首创性地采用数据挖掘方法，采用真实的网络借贷微观大样本数据，特别是用户“行动轨迹”信息和“社会资本”信息的挖掘，探究影响网络借贷用户是否会再借款的行为模式，以及所揭示出来的信任信号，以提高借贷成功率和投资者效率，为提高网络借贷平台运营效率创造价值。

三、研究方法及数据选择

（一）研究方法及研究思路

评分模型的开发，一般从可解释性、模型预测的准确性、模型的稳健性三个重要层面去思考。最常用的算法当属Logistic回归，其原理类似于线性回归，优点在于可解释性强，模型简单，

不足之处是精度一般。人工智能的相关算法也越来越多地用于评分模型的开发，比如决策树、Ensemble Method、SVM、神经网络等，但是这些算法都有自己很强的使用场景。相对于Logistic回归，决策树的精度有所提高；基于决策树基础之上的Ensemble Method，其开发的模型鲁棒性更好；SVM一般适用于小数据集，易过度拟合；神经网络则是一个完全的黑箱操作，可解释性较差，而且一般适用于超大数据集。

XGBoost算法采用分布式加载数据、分布式训练数据，同时对损失函数做了二阶的泰勒展开，并在目标函数之外加入正则项整体求最优解，用以权衡目标函数的下降和模型复杂程度，避免过度拟合。XGBoost算法是在GBDT算法基础之上改进的，是通过构建多棵树融合的模型，属于Ensemble Method的一种。Ensemble Methods的目标在于将几种机器学习的算法结合或者把一种算法的不同参数组合，该算法鲁棒性非常好，构建的模型也是基于树模型的基础之上，可解释性较Logistic回归稍差，但是不像神经网络完全的黑箱操作。XGBoost算法也属于Gradient Boost的一种，与传统Boost的区别是，每一次的计算是为了减少上一次的残差（residual），而为了消除残差，我们可以在残差减少的梯度（Gradient）方向上建立一个新的模型。在Gradient Boost中，每个新模型的建立是为了使之前模型的残差往梯度方向减少，这与传统Boost对正确、错误的样本进行加权有着很大的区别。

本文针对网络借贷中的借款人数据进行数据清洗、特征工程、去共线性，算法选用鲁棒性非常好的XGBoost，结合CV进行十折交叉验证，再通过PSO算法优化参数，以挖掘网络借贷中再次借款人所具有的特征，为平台挖掘优质用户，特别是挖掘再次借款人的行为数据及其在平台中的交互活动产生的社会关系交互数据，通过网贷平台上的行动轨迹以揭示其再次借款的信号。最后，基于模型结果，对是否再借款的信息特征进行重要性排序，依据模型的打分，并给出具体的可视化展示。

（二）数据来源与指标说明

本文数据来源于融360平台提供的26 000个用户的脱敏数据，其中申请再借款的用户个数与未申请再借款的用户个数分别为12 859个和13 141个，本文目标变量为用户是否再借款。融360是我国最大的网络贷款平台，平台的一端是有借款需求的个人消费者和小微企业，另一端是有投资资金的金融机构及其提供的数百万种金融产品，包括银行、小贷、担保、典当等。融360通过搜索和推荐服务来撮合借款和贷款。通常，借款用户进入平台后，会通过搜索和推荐服务找到合适的贷款产品，填写自己的个人基本资料并提交贷款订单。金融机构在平台收到订单后，对用户资质进行风控审核，最终决定是否通过用户的订单。

详细的变量信息说明见表1，其中包括基本信息、消费信息、行为标签和社交关系信息：（1）用户的基本信息数据：用户在融360平台上提交的基本资料，如年龄、性别、职业、教育信息等，包含用户在平台上多次修改个人信息的记录，共253 006条；（2）用户的行为轨迹数据：用户在融360平台的行为轨迹标签，共687 374条；（3）用户的借款行为信息数据：用户在融360平台上提交的借款申请，包括申请了什么产品、贷款额度及还款情况等记录，共233 450条；（4）用户的信用卡消费行为信息数据：用户向融360平台上申请的信用卡及信用卡消费行为的信息数据，一个用户一般有多条信用卡消费记录，共677 540条；（5）用户的社会交往信息：融360为用户提供了社交机制，用户的“朋友数目”越多代表社会资本较多。用户在平台上的社交关系及关系强度通过两个指标来揭示，用户社交关系1，是用户与其他用户的关注情况，共7 367 375条；用户社交关系2，是用户与其他用户的深层次关系情况，共1 989 230条。

表 1　数据说明

说　明	变量名称	变量含义	说　明	变量名称	变量含义
用户基本信息	user_id	用户ID	用户信用卡消费行为信息	user_id	用户编号
	age	年龄		prior_period_bill_amt	上期账单金额
	sex	性别		credit_lmt_amt	信用卡额度
	education	教育程度		current_bill_bal	本期账单余额
	live_info	房屋类型		is_cheat_bill	是否恶意账单
	occupation	职业类型		current_bill_amt	本期账单金额
	marital_status	婚姻状态		circle_interest	循环利息
	local_hk	户口类型		nadd_jifen	新增积分
	salary	薪水		avlb_bal_usd	可用余额美元
	school_type	学校类型		card_type	卡类型
	company_type	公司类型		credit_lmt_amt_usd	信用额度美元
	flow	月盈利余额		curr	币种
	business_type	营业类别		current_min_repay_amt_usd	本期最低还款额美元
	personnel_num	公司规模		current_convert_jifen	本期兑换积分
	gross_profit	毛利润率		bill_id	记录编号
	business_year	营业年限		prior_period_repay_amt	上期还款金额
	pay_type	收入类型		curt_jifen	当前积分
用户社交关系1（直接关注情况）	user_1	用户1		current_bill_min_repay_amt	本期账单最低还款额
	user_2	用户2		cost_cnt	可消费笔数
用户社交关系2（深层关系情况）	user_1	用户1		adj_amt	调整金额
	user_2	用户2		prior_period_jifen_bal	上期积分余额
	relation2_type	关系类型		current_adj_jifen	本期调整积分
	relation2_weight	关系权重		avlb_bal	可用余额
	time	时间		pre_borrow_cash_amt_usd	预借现金额度美元
用户借款行为信息	product_id	产品ID		pre_borrow_cash_amt	预借现金额度
	expect_quota	申请金额		repay_stat	还款状态
	money_function	贷款用途		current_repay_amt_usd	本期应还款金额美元
	max_month_repay	最大月还款		current_award_jifen	本期奖励积分
	recode_time	记录时间	用户行为标签	user_id	用户ID
				rong_tag	应用ID

表1所列的信息特征基本都是一个用户存在多条记录，总共信息维度为55维，其中user_id为用户唯一标识。我们需要从54个维度提取出特征，这些特征要能说明用户是否再借款。

（三）特征工程

特征工程是将原始数据转化为特征，准备和选择的特征越好，模型预测的结果会越好。本文进行特征工程采用的原则是：尽可能地避免清洗数据，尽量避免构建特征时的信息损失。下面说明对以上信息表进行数据清洗以及构建特征工程的方法。

针对用户基本信息、用户借款和信用卡消费信息，利用用户ID（user_id）进行分组后统计各指标的相关统计量，包括最小值、最大值、求和、均值、标准差、频数。针对用户社交关系，将原始的用户对应关系转换为图，并用图算法，深入统计用户的“社会资本”关系数据，挖掘第一层直接关注为好朋友的数量，以及第二层间接关注为好朋友的数量，通过程序测试，最终统计了其频数。最终本文构建的特征工程维度为251维，包含用户标识变量user_id。

在构建以上特征工程之后，需要进一步剔除具有高相关性的特征，为后续建模提供更高质量的数据集。Logistic回归模型对特征共线性比较敏感，但XGBoost算法对共线性处理没有那么强的要求，一般考虑剔除特征工程后存在的相关系数为1的变量。本文通过构建特征工程后的维度为251，去除相关系数为1的变量后，维度为193。检查数据后发现，存在此现象的原因在于数据集的稀疏性。

四、实证分析

（一）建模步骤

为了更好地模拟实际场景，需要解决两个问题：第一，构建一个尽可能准确的模型以预测用户是否会再借款；第二，在模型准确性的基础上，提取出有意义的特征，对用户的再借款特征进行挖掘和分析。为了达到以上两个目的，本文构建了大量的特征工程，再剔除一部分高相关的特征，通过XGBoost算法完成特征寻找的工作，为接下来构建用户再借款预测模型提供基础。

本文通过以下两个步骤构建模型：第一，为了避免过拟合现象找到模型的最优参数组合，采用CV进行十折交叉验证的方式，通过PSO算法搜索得到XGBoost模型最优参数，其中，对13个参数进行设定，调整的参数包括7个，设置保持不变的参数包括6个，最终得到模型的最优参数见表2；第二，将数据集按照7:3的方式，拆分为训练集和测试集。

表 2　模型的参数设置

	参数	参数值	参数说明
参数设置后固定不变	booster	gbtree	使用基于树的模型进行提升计算
	objective	binary：logistic	解决二分类的逻辑回归
	eval_metric	AUC	每次检验的评价指标
	scale_pos_weight	1.2	正负样本的权重比
	missing	–1	数据缺失说明
	Seed	520	每次训练抽样的随机种子
需要调整的参数	max_depth	9	每棵树的最大深度
	gamma	15	后剪枝参数
	eta	0.045	提升计算过程中的收缩步长
	lambda	17	L2 正则的惩罚系数
	subsample	0.89	用于训练模型的子样本占整个样本集合的比例
	colsample_bytree	0.89	建立树时对特征采样的比例
	num_round	16	构建的树个数

表2中的参数num_round是迭代的次数，也是最终构建的树的个数。一般选择控制比较小的num_round，再调整其他参数，在num_round比较小的时候，是便于模型解释的，所以XGBoost算法的可解释性也很高。

（二）模型结果

模型的准确性评价指标很多，本文采用最常用的指标AUC值来衡量模型的好坏。本文构建的再借款模型的训练集AUC结果为0.711 2，测试集的AUC结果为0.698 8。是否再借款的预测难度很大，AUC已达到0.7，表明该模型已能为网贷平台提供很好的预测和决策。

接下来，本文进一步挖掘潜在的用户特征，即具有怎样特征的用户会选择再借款。通过XGBoost算法，最终进入模型的特征个数为63个，我们进一步给出是否再借款预测模型的特征重要性。

依据特征重要性评分排序，图1绘制了特征重要性的散点图，横轴为各特征，纵轴为评分，曲线score_feature代表的是模型给出的变量重要性评分。

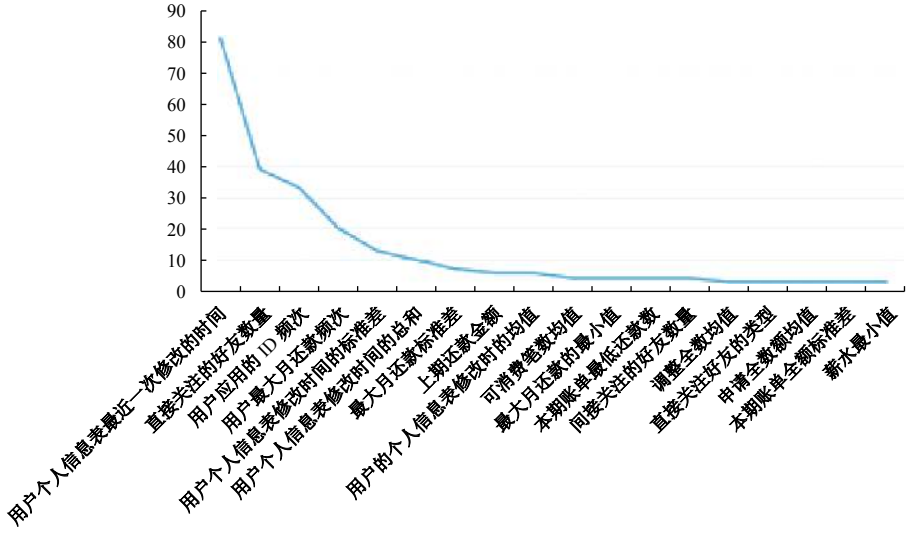


图 1　XGBoost模型的特征重要性散点图

图1显示了进入模型的特征评分，评分呈现急剧下降，表明能反映目标的特征很少，也说明预测用户是否会再借款的难度很大。按重要性评分排在前18的特征主要可以归为以下四类：第一类为用户在平台上对个人信息修改的“行动轨迹”记录数据，包括排序在第1、3、5、6和9的特征。第二类为用户在平台上的社会交往网络记录信息，包括排序在第2、13和15的特征。第三类为用户在平台上申请贷款历史的“行动轨迹”数据，包括排序在第4、7和15的特征。第四类为用户的消费记录的“行动轨迹”数据信息，包括排序在第8、10、12、14和17的特征。以上信息皆为用户在网站上留下的“行动轨迹”数据和“社会交往”数据，借款人自身的住房、学历、性别等“硬信息”在“是否再借款”预测模型中并没有起到预测作用，仅“工资”特征排在重要性评分的第18项。可能的原因是，本文针对已成功申请过借款的用户是否“再借款”进行预测，对于已成功申请过借款的用户来说，其“硬信息”已经获得网贷平台和贷款人的认可，对于用户“是否再借款”显得不是那么重要了。

进一步地，针对重要性排名前五的特征进行描述性统计分析。如表3所示，与用户再借款关系非常密切的前五个特征分别是：用户的个人信息表最近一次修改的时间、用户第一层朋友个数、用户应用ID频次、用户最大月还款频次、用户的个人信息表修改时间的标准差。

表 3　模型特征重要性排名前五变量的描述性统计分析

Feature	max_tm_encode	1u_friends	tag_count	max_month_repay_count	std_tm_encode
特征说明	用户的个人信息表最近一次修改的时间	直接关注的好友数量	用户应用ID频次	用户最大月还款频次	用户的个人信息表修改时间的标准差
Score_feature	81	39	33	20	13
最小值	4 804 428	–1	–1	0	48.853 86
最大值	21 856 564	18 638	306	21	11 096 288
均值	17 849 921	204.048 7	18.942 5	0.559 65	1 295 321

续表 3 模型特征重要性排名前五变量的描述性统计分析

Feature	max_tm_encode	1u_friends	tag_count	max_month_repay_count	std_tm_encode
特征说明	用户的个人信息表最近一次修改的时间	直接关注的好友数量	用户应用ID频次	用户最大月还款频次	用户的个人信息表修改时间的标准差
标准差	4 469 699	421.917 9	27.684 8	1.408 32	2 127 014
缺失个数	0	2 959	13 519	0	0
25分位数	16 713 646	50	–1	0	585.646 6
50分位数	19 672 730	120	–1	0	48 975.17
75分位数	21 157 891	238	34	0	1 963 267

注：原始特征中缺失用–1标识。

（三）特征结果分析

通过构建XGBoost模型，我们提取了前五个与再借款行为最相关的特征，为了辅助我们理解这些特征，我们借助可视化的方法以展示这些特征与用户是否会再借款之间的关系。需要说明的是，原始总样本中未申请再借款与申请再借款的比例为1.022:1。

1. 第一行为特征变量“用户的个人信息表最近一次修改被记录的时间”

模型中最显著的特征是max_tm_encode，该特征是指“用户的个人信息表最近一次修改的时间”，该值越大表示时间越新，越是“近期”。如图2所示，横轴表示该特征值的百分位区间，分位数越大表示时间越“新”，表示用户越在近期修改了个人信息，纵轴是用户数量。可以看出，max_tm_encode在20分位点是个明显的分割点，在低于20分位点以下时，未申请再借款人数与申请再借款的比例大约为4∶1，max_tm_encode在“大于20分位数”时，未申请再借款人数明显小于申请再借款人数。这表明用户越久未修改自己的个人信息表，越倾向于不选择再借款，在最近修改自己的个人信息的用户越有可能申请再借款。这也意味着，在网贷平台上用户越活跃，越在乎平台上个人信息的准确性，越有可能会再次申请借款。

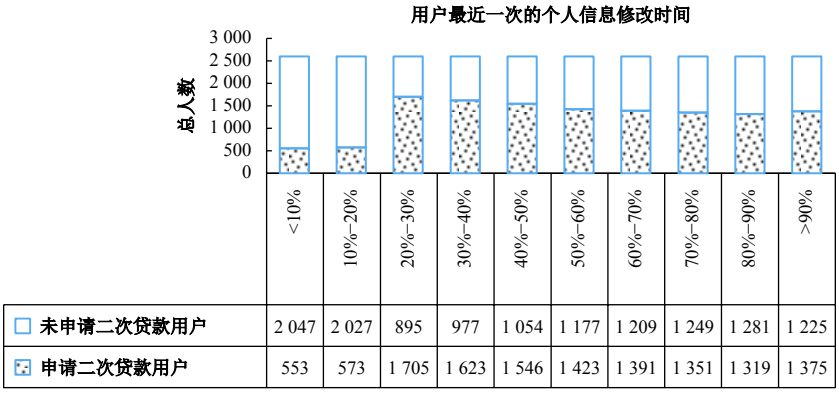


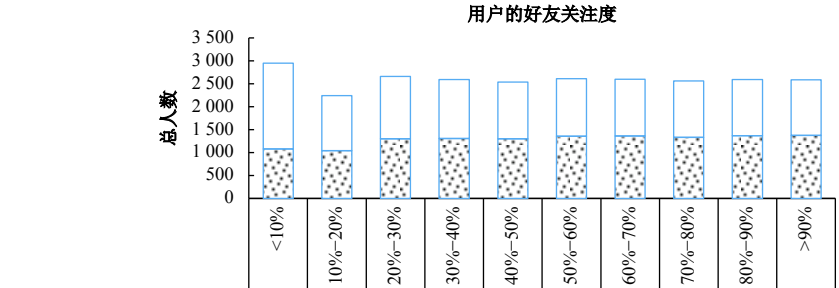
图 2 用户最近一次个人信息修改时间分位数与用户是否再借款统计图

注：横轴坐标表示各分位点，例如“10%–20%”表示被分析变量取值处于10至20分位点之间时未申请再借款用户与申请再借款用户的数量关系。以下各图坐标意义相似，不再赘述。

2. 第二社会交往特征变量“用户的好友关注度”

用户的好友关注度是指网络借贷平台上与用户有直接关注关系的好友数量，变量名为1u_friends。如图3所示，横轴表示“直接关注的好友数量”的分位点区间，可以看出，数量总额在“小于10分位数”时，未申请再借款与申请再借款的比例大约为1.73:1，而数量总额在“大于90分位

数”时，未申请再借款与申请再借款的比例大约为0.88:1。相对于原始数据中未申请再借款与申请再借款的比例1.022:1而言，表现出明显的反差。该趋势表明，倾向于再借款的用户，拥有越多直接相关的朋友。用户的好友关注度指标，向网络借贷平台传递一种正能量，即社会资本越多的用户越倾向于再次申请借款。

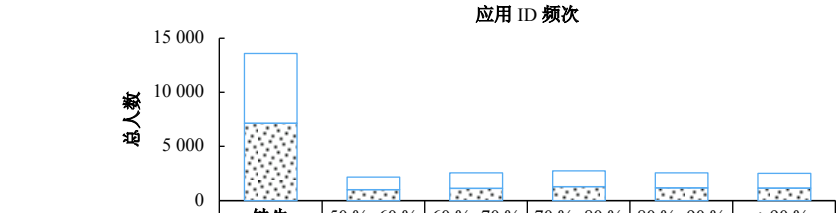


	<10%	10%–20%	20%–30%	30%–40%	40%–50%	50%–60%	60%–70%	70%–80%	80%–90%	>90%
未申请二次贷款用户	1 875	1 199	1 365	1 292	1 239	1 258	1 239	1 233	1 227	1 214
申请二次贷款用户	1 084	1 047	1 301	1 307	1 302	1 361	1 367	1 336	1 374	1 380

图 3　用户的好友关注度与用户再借款分位数统计图

3. 第三行为特征变量“应用ID频次”

“tag”是融360公司提供的一个组合指标，中文解释是“应用ID”，表3中的变量名“tag_count”指的就是用户应用ID出现的频次。该频次存在大量的缺失，分位数在前50都是缺失，未申请再借款与申请再借款的比例大约为0.9:1；在后50分位数中，未申请再借款与申请再借款的比例大约为1.17:1，表明应用ID的频次越高，再次申请借款的可能性越高。这意味着，在网络借贷平台上越活跃的用户，其再次申请借款的概率越大。参见图4。

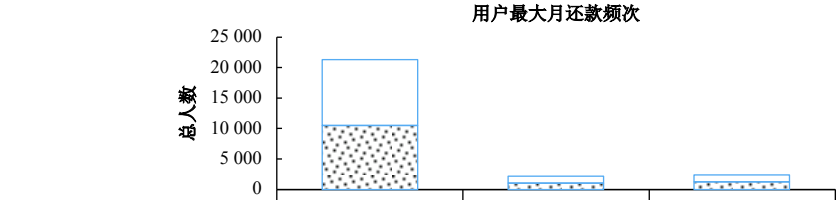


	缺失	50 %–60 %	60 %–70 %	70 %–80 %	80 %–90 %	>90 %
未申请二次贷款用户	6 406	1 159	1 409	1 433	1 383	1 351
申请二次贷款用户	7 113	1 002	1 132	1 292	1 172	1 148

图 4　应用ID频次与用户再借款分位数统计图

4. 第四行为特征变量“用户最大月还款频次”

“max_month_repay”是“用户最大月还款”。变量“max_month_repay_count”指的是用户最大月还款出现的频次。该频次存在大量的0，分位数在前80值都是0，未申请再借款与申请再借款的比例大约为1.03:1，非常接近总样本中未申请再借款与申请再借款的比例为1.022:1。而在大于90分位数时，未申请再借款与申请再借款的比例大约为0.932:1，说明用户最大月还款的频次越多，越倾向于申请再借款。用户欲再次借款，为了保持良好的信用记录，按时还款的同时，一般还款的频次也会较多，这些行为是为了进一步向借贷平台或投资者揭示其良好的信用。参见图5。

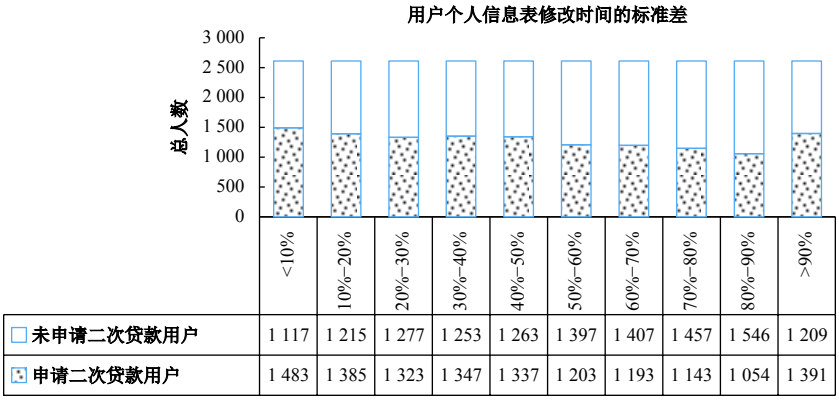


	0	80 %–90 %	>90 %
未申请二次贷款用户	10 849	1 124	1 168
申请二次贷款用户	10 535	1 071	1 253

图 5　用户最大月还款频次与用户再借款分位数统计图

5. 第五行为特征变量“用户个人信息表修改时间的标准差”

用户修改个人信息表时间的标准差，反应了用户修改个人信息表时间的波动情况。标准差越小则数据的波动越小，表明客户集中在某个时间段修改自己的个人信息。从图6可以看出，在低于50分位数时，申请再借款的人数比例更高，在高于50分位数之后（90分位数除外），申请再借款人数比例下降，表明申请再借款的客户修改个人信息表时间的标准差普遍偏小，集中在某个时间段修改自己的个人信息。比较特殊的是，在标准差大于90分位数时，申请再借款的人数比例却较高，表明这些用户一直在关注网贷平台，维护平台上的个人信息。



	<10%	10%–20%	20%–30%	30%–40%	40%–50%	50%–60%	60%–70%	70%–80%	80%–90%	>90%
未申请二次贷款用户	1 117	1 215	1 277	1 253	1 263	1 397	1 407	1 457	1 546	1 209
申请二次贷款用户	1 483	1 385	1 323	1 347	1 337	1 203	1 193	1 143	1 054	1 391

图 6　用户个人信息表修改时间的标准差与用户再借款分位数统计图

综上所述，网络借贷平台用户的行为轨迹数据，可以归结为两类：一类是为了减少信息不对称（保持良好信用）的行为，如还贷款频次，修改个人信息的时间及其标准差，用户为了向投资者揭示自己良好的信用而维护自己的信息和声誉；另一类是平台上的社会网络层面，通过维护平台上的社会网络以此构建自己的社会资本，在网络借贷平台上的朋友关系良好，目的也是为了建立和揭示信任。因此，网络借贷平台通过用户的行为轨迹，特别是能揭示其信任信号的行为信息和平台社会网络交流信息，可预测出用户是否会再借款。

五、结论与建议

网络借贷通过互联网模式撮合了被排除在传统银行借贷之外的个人或小微企业之间的借贷，平台自身积累了用户大量的基本信息，如用户借贷和消费的“行为轨迹”大数据，以及代表“社会资本”的社交网络大数据。针对网络借贷中“高频次”、“反复借贷”的特征，本文通过数据

挖掘技术建立了再借款评估模型，以分析网络借贷中再次申请借款的用户具有怎样的行为轨迹和社交特征，由此挖掘平台的优质和忠诚用户，以帮助投资者准确判断借款人，从而降低借贷风险，提高平台的运营效率，也便于平台为用户提供更好的金融服务。具体而言，本文采用鲁棒性非常好的XGBoost算法，利用网络借贷平台2.6万个用户68万多条行为轨迹数据和社交网络数据，首创性地建立了“用户是否再贷款”预测模型，并对关键特征进行可视化分析。得出以下主要结论：(1)通过XGBoost算法构建模型，并进行十折交叉验证，再通过PSO算法搜索最优参数，使得最终模型的训练集AUC达到0.71。测试集的AUC将近0.7，表明本文建立的模型能为网络借贷平台针对用户“是否会再次申请借款”提供很好的预测。(2)网络借贷平台中用户是否会申请再借款，与用户的个人基本信息（“硬信息”）的关系并不是很强，反而是用户在平台上的借贷和消费的“行为轨迹”信息，以及在平台上的社会交往行为的社会资本类信息，更具有信任信号的揭示作用，预测能力更强。(3)在网络借贷平台上积极维护个人信息、保持自己良好的还款记录的用户，再次申请借款的可能性很高。用户最近一次修改个人信息的时间越近、用户应用ID频次越高、用户最大月还款的频次越高，越有可能申请再借款，成为网贷平台上的忠实用户。这也意味着，在网络借贷中想要多次申请借款的用户更须注重建立和保持自己良好的信用，揭示信任信号，以便在多次借贷博弈中成本最小化。(4)在网络借贷平台上关注的朋友数量越多，具有良好社会资本，积极维护自己社会网络关系的用户，再次申请贷款的可能性也更高。借款人通过平台的社交网络关系构建社会资本，向投资者揭示自身的声誉信号，这类用户往往再次申请借款，成为平台的优质客户。

综上所述，网络借贷平台可借助自身的互联网和大数据优势，通过强化借款人的约束机制、放大借款人的声誉机制，通过数据挖掘以降低信息不对称，提高交易效率。一方面，通过构建更丰富的社会网络交流平台，或者直接与大型社交网络（例如微信、QQ等）相联通，以获取用户的社交网络大数据，通过社会资本数据揭示信任信号；另一方面，网络借贷平台应该加大信息的有效挖掘和披露，通过挖掘用户潜在的行为轨迹和社会网络关系，利用互联网渠道和大数据技术来降低信息不对称，降低投资者的搜寻成本，不仅可提高交易效率还可以吸引更多的投资者。

行业协会可致力于各个网络平台打破信息孤岛，建立网络借贷的黑名单制度，构建互联网金融信用体系，加大对失信借款人的处罚力度和约束机制。这都有助于抑制借款人的短期投机行为，增强借款人维护自身声誉的动机，放大声誉的激励机制，防止网贷市场出现“劣币驱逐良币”现角，促进网络借贷市场的健康良性发展。

主要参考文献：

[1] 蒋翠清，王睿雅，丁勇. 融入软信息的P2P网络借贷违约预测方法[J]. 中国管理科学，2017，(11).
[2] 李焰，高弋君，李珍妮，等. 借款人描述性信息对投资人决策的影响——基于P2P网络借贷平台的分析[J]. 经济研究，2014，(S1).
[3] 廖理，李梦然，王正位. 聪明的投资者：非完全市场化利率与风险识别——来自P2P网络借贷的证据[J]. 经济研究，2014，(7).
[4] 莫易娴. P2P网络借贷国内外理论与实践研究文献综述[J]. 金融理论与实践，2011，(12).
[5] 孙武军，樊小莹. 从业经历和教育背景是否能提高借贷成功率？——来自P2P平台的经验证据[J]. 中央财经大学学报，2016，(3).
[6] 王会娟，何琳. 借款描述对P2P网络借贷行为影响的实证研究[J]. 金融经济学研究，2015，(1).
[7] 王会娟，廖理. 中国P2P网络借贷平台信用认证机制研究——来自“人人贷”的经验证据[J]. 中国工业经济，2014，(4).
[8] 王硕. 国内P2P网络借贷平台运营模式分析[J]. 当代经济，2015，(22).
[9] 王修华，孟路，欧阳辉. P2P网络借贷问题平台特征分析及投资者识别——来自222家平台的证据[J]. 财贸

经济，2016，(12).
[10] 杨振能. P2P网络借贷平台经营行为的法律分析与监管研究[J]. 金融监管研究，2014，(11).
[11] 张国文. 论P2P网络借贷平台的风险防范与监管[J]. 武汉金融，2014，(4).
[12] Davis K E，Gelpern A. Peer-to-peer financing for development：Regulating the intermediaries[R]. NYU Law and Economics Research Paper No.10–22，2010.
[13] Freedman B S，Jin G Z. Do social networks solve information problems for peer-to-peer lending？ Evidence from prosper.com[R]. NET Institute Working Paper No.08–43，2008.
[14] Greiner M E，Wang H. The role of social capital in people-to-people lending marketplaces[A]. Proceedings of 2009 international conference on information systems[C]. Phoenix，Arizona，USA：ICIS，2009.
[15] Herzenstein M，Andrews R L，Dholakia U M. The democratization of personal consumer loans？ Determinants of success in online peer-to-peer lending communities[R]. SSRN Working Paper No. 1147856，2008.
[16] Herzenstein M，Sonenshein S，Dholakia U M. Tell me a good story and i may lend you money：The role of narratives in peer-to-peer lending decisions[J]. Journal of Marketing Research，2011，48(SPL)：S138–S149.
[17] Iyer R，Khwaja A I，Luttmer E F P，et al. Screening in new credit markets：Can individual lenders infer borrower creditworthiness in peer-to-peer lending？ [R]. AFA 2011 Denver Meetings Paper，2009.
[18] Klafft M. Online peer-to-peer lending：A lenders' perspective[A]. Proceedings of the international conference on E-learning，E-business，enterprise information systems，and E-government[C]. Las VegasCSREA Press，2008：371–375.
[19] Larrimore L，Jiang L，Larrimore J，et al. Peer to peer lending：The relationship between language features，trustworthiness，and persuasion success[J]. Journal of Applied Communication Research，2011，39(1)：19–37.
[20] Lin M F，Prabhala N R，Viswanathan S. Judging borrowers by the company they keep：Friendship networks and information asymmetry in online peer-to-peer lending[J]. Management Science，2013，59(1)：17–35.
[21] Pope D G，Sydnor J R. What's in a Picture？ Evidence of discrimination from prosper.com[J]. Journal of Human Resources，2011，46(1)：53–92.
[22] Slattery P. Square pegs in a round hole：SEC regulation of online peer-to-peer lending and the CFPB alternative[J]. Yale Journal on Regulation，2013，30：233–275.
[23] Weiss G N F，Pelger K，Horsch A. Mitigating adverse selection in P2P lending: Empirical evidence from prosper.com[R]. Social Science Electronic Publishing，2010.
[24] Yum H，Lee B，Chae M. From the wisdom of crowds to my own judgment in microfinance through online peer-to-peer lending platforms[J]. Electronic Commerce Research and Applications，2012，11(5)：469–483.

Can Behavioral Track and Social Information Predict Whether Online Lending Users will Refinance? An Empirical Analysis Based on Online Lending

Huang Jing, Miao Shilei

(*Business College*, *Shanghai Normal University*, *Shanghai* 200234, *China*)

Summary: Online lending has the characteristics of "high frequency" and "repeated borrowing". Users often have the demand of refinance, so how to use the information accumulated by the online platform, especially users' "Behavioral Track" data and "Social Information" data, to dig the high quality and loyal users who have the demand of refinance is very important, which can not only improve the operational efficiency of the online lending platform, but also keep the online lending market stable in the long-term. In this paper, with the XGBoost method, we use the information of 26,000 users from the online lending platform to

create a refinance forecasting model. The main conclusions are as follows: (1)To predict whether users refinance in online lending, "hard information" (or personal basic information) provided by users has no signal effect. "Trajectory" information and "social interaction" information are more revealing of the trust signal. (2)Users who actively maintain their personal information and keep good repayment records on the online lending platform are highly likely to apply for loans again. The closer the time when users last modified their personal information, the higher the frequency of user ID application and maximum monthly repayment, the more likely they are to apply for refinance and become loyal users on the online lending platform. (3)Users who follow more friends on the online lending platform, have good social capital and actively maintain their social network relationships, are more likely to apply for loans again. Borrowers build social capital through social network relations of the platform and reveal their own reputation signals to investors. Such users will often refinance and become high-quality customers of the online lending platform. Therefore, the online lending platform can make use of its Internet and big data advantages to strengthen borrowers' constraint mechanism, enlarge their reputation mechanism, and reduce information asymmetry to improve the online transaction efficiency through data mining.

Key words: online lending; behavioral track; social interaction; trust signal reveal

(责任编辑：王西民)

(上接第92页)

migrants. In the face of possible endogenous problems, this paper uses the 1986 Compulsory Education Law to promulgate this quasi-natural experiment construction tool variable, and adopts the 2SLS method for robust regression analysis. The empirical results show that the increase in the number of years of education does help the township immigrants identify with the city. On the basis of the basic conclusions, this paper also carries out corresponding channel analysis, and finds that education plays an active role in the process of urban migrants obtaining urban household registration, urban housing and married urban residents, and thus promotes the urban integration of immigrants. This study has important policy implications: Under the background of the urbanization of population, it is of great significance to continuously promote the development of rural education and strive to improve the education level of immigrants in rural areas. Based on the research conclusions, the following policy recommendations are proposed: First, gradually correct the urban bias of education expenditure, establish a long-term supply mechanism to guarantee rural education funds, narrow the gap between urban and rural education investment, and promote the rapid and steady development of rural education. Second, vigorously develop secondary education in rural areas, especially medium and high vocational education, and increase the educational opportunities and education level of rural residents. Third, actively formulate and improve education support policies for low-income families in rural areas and the scholarships and student loans for poor students, and create conditions for rural poor students to continue their studies.

Key words: educational level; rural-urban migrants; urban integration; urban self-identity

(责任编辑：王西民)