网络借贷违约风险分析——基于数据挖掘

钟教聪,方 华
(上海理工大学,上海 200093)

摘 要:以P2P网络借贷为例,从人人贷中选取2015—2018年共7 559条记录,通过数据挖掘模型来对借款人违约风险进行分析,并识别出影响借款人违约的主要因素,这些数据挖掘模型主要包括决策树、支持向量机和随机森林。主要结论包括:第一,运用数据挖掘模型来预测违约风险效果都很好,其中最好的是随机森林;第二,特征重要性程度前五依次为信用等级、借款金额、借款周期、借款利率、借款人所在企业的规模。

关键词:P2P网络借贷;数据挖掘;违约风险
中图分类号:F832 文献标志码:A 文章编号:1673-291X(2020)10-0088-04

引言

近年来,互联网与金融的结合更加广泛,互联网金融凭借其支付优势、流程优势等优势逐渐深入人心,同时也对我国经济的发展起到了很大的促进作用。其中,P2P网络借贷是互联网金融的一个分支。P2P网络借贷,通常是指个体和个体通过互联网平台进行的直接借贷活动。艾瑞咨询统计结果显示,截至2017年,我国网络借贷超过了2万亿元,且年增长率高达40%,用户高达2亿人,相较2016年增长23.1%,可见网络借贷发展之蓬勃。

P2P网络借贷开始出现是在英国,因为其相较于传统银行更加方便,回报率高,很快便快速蔓延至其他国家。2007年6月,我国第一家P2P网贷公司成立,从此网络借贷在我国拉开了序幕。在2013年前,我国P2P网贷平台发展的很慢,属于萌芽期。2013开始,我国P2P网贷行业在用户和平台都开始爆发性增长。但是在爆发性增长的同时也伴随着很多风险,截至2017年,停业的P2P网贷平台已达1 500家,网贷平台坏账率普遍达到了10%以上,这显著高于传统金融机构。网贷平台的高风险,有一个主要原因是,网贷不需要抵押,借款人违约成本较低,如果出现很多借款人违约,则会对平台现金流产生影响,会影响平台的可持续发展。在此背景下,对借款者的违约风险进行分析显得尤为重要。

本文主要运用数据挖掘的方法,基于数据借款人信息,找出影响借款人的违约因素,以期能给网贷平台和投资者提供些参考。本文选用的模型相对于传统的风险分析模型主要优势是,传统的模型大多需要设定参数,对前提假设有很严格的限制,如最小二乘模型要求数据必须符合正态分布、序列没有关联且没有噪声。Logistic要求自变量不能存在多重共线性,而数据挖掘对数据并无限制。

一、文献综述

由于网络借贷的快速发展,对金融业产生了较大的冲击,因此引起了学术界的广泛关注,中外学者对进行了很多关于网络借贷违约的研究。

从违约风险来看。由于信息不对称使得投资人和网贷平台不能很好地评价借款人违约风险的大小,从而增加了投资者和网贷平台的风险(刘丽丽,2013)。同时由于网贷借款人在借款人并不是抵押借款,违约成本比较低,且贷款用途没有限制,这使得贷款风险显著增加(李渊琦、陈芳,2015)。社会资本的存在能有效降低借款人的违约风险,这些社会资本包括借款列表被推荐的额次数、是否加入小组、增加投资者中朋友的个数等等(缪莲英、陈金龙,2014)。通过使用多元线性回归模型对拍拍贷进行违约风险分析,发现随着年龄的增加违约风险越低(刘鹏翔,2017)。借款人声誉能有效缓解信息不对称,声誉变量包括借款人以往违约次数和借款成功次数为代表,实证得出借款人声誉对违约风险的识别效应,且如果借款人还款能力增加,这种识别能力也会增强(李鑫,2019)。以拍拍贷为例,研究学历在网络借贷上的作用,发现随着学历的上升,借款人逾期的风险越小,且借款成功的概率更高(程瑶,2018)。

就研究模型来看,经典的预测借款人违约的模型,如Logistic、Probit,OLS预测效果有太多的约束,如对样本要求比较严格,在特征较为复杂的情况下,预测效果会大打折扣(Hill Griffiths and Lim,2011)。相较于经典的预测模型,数据挖掘模型对样本没有较多约束,且能应对更为复杂的自变量,通常情况下,预测效果好于经典预测模型(Goyal,A.and R.Kaur,2016)。

收稿日期:2019-10-08
作者简介:钟教聪(1995-),男,海南昌江人,硕士研究生,从事互联网金融研究。

二、模型选择

本文所选用的数据挖掘模型包括支持向量机(SVM)、决策树(DT)和随机森林(RF),这三种模型都是监督学习算法,都是可以通过训练样本获得最优模型的。

(一)支持向量机

支持向量机的目标是创建一个平面边界,称为超平面,从而将具有不同性质的样本进行划分,划分的原则是间隔最大化。支持向量机从20世纪90年代开始快速发展,目前在很多领域都得到广泛应用。支持向量机可以将低维度空间样本分类的问题投影到高维度空间,从而可以在新的空间上得出最优超平面。

目前,支持向量机模型常用来解决分类问题的核函数包括以下四种:线性核函数、多项式核函数、S形核函数以及高斯RBF核函数。本文所采用的是线性核函数,其设定如下:

$$K(\vec{x_i},\vec{x_j})=\vec{x_i}|\vec{x_j}$$

其中,x_i表示第i个特征。

(二)决策树

决策树是一种有监督的算法,按照一定的划分规则,对数据进行持续的划分,最后根据根据投票结果进行分类。决策树在任何领域上几乎都能用,可以说是应用最广泛的数据挖掘模型了。决策树的算法主要包括ID3算法、GART算法和C4.5算法,本文采用的是C4.5算法。

C4.5算法用信息增益率选择决策属性。C4.5算法有两个步骤,第一,先选取一个属性A_i,按照A_i的某个值将n维空间进行划分成两个部分。第二,按照第一步重新选择另一个属性进行划分,直到n维空间都被划分了。C4.5算法划分的标准是信息增益率(Info Gain Ratio)指标。假设数据集D有m个类别,数据D的熵可定义为:

$$Info(D)=-\sum_{i}^{m}p_i log_2 p_i$$

其中,p_i表示类别i占样本的比率,数据集的种类越多,则Info(D)越大,当数据集只有一个类别时,Info(D)为0。现假定属性A对数据集D进行划分,且划分为K个D_j子集,则划分后的数据集D的熵为:

$$Info_A(D)=\sum_{j=1}^{K}\frac{|D_j|}{|D|}Info(D_j)$$

$|D|$表示数据集D的样本量,$|D_j|$表示数据集D_j的样本量,则信息增益为:

$$Gain(A)=Info(D)-Info_A(D)$$

要想得出信息增益率,必须先求出使用“分裂信息”值,分裂信息定义为:

$$SplitInfo_A(D)=-\sum_{j=1}^{k}\frac{|D_j|}{|D|}log_2\left(\frac{|D_j|}{|D|}\right)$$

在C4.5算法中,信息增益率最大的属性为划分标准。最后,信息增益率为:

$$GainRatio(A)=\frac{Gain(A)}{SolitInfo_A(D)}$$

(三)随机森林

随机森林是一种集合学习的方法,随机森林通过随机建立一个森林,这森林里包括很多个决策树,随机森林里的每棵决策树都是相互独立的。在建立随机森立后,当输入一个样本,则随机森林里的每棵决策树树都会对样本进行决策,然后在通过这些决策树进行投票,从而得出最终的预测值。随机森林有效地提高了预测精度,并且能够给出每个特征变量的重要程度。

三、实证分析

第一,数据来源。人人贷是我国较早进行网络借贷的平台,也是发展的比较好的平台。本文通过python爬虫的方法从人人贷平台上选取了2015—2018年上半年的个人借款数据,由于存在到大量与个人信用无关的信息,如借款人昵称、贷款编号等,若加入模型,可能会造成不必要的干扰。此外,有些变量是字符型的,也改成数值型。最终,借款人的信息包括个人信息(借款人年龄、学历、性别、工作区域、是否有房、是否有车、是否有房贷、是否有车贷,婚姻状况、工资、公司规模以及工龄);借款人信用情况(信用评级);借款信息(借款利率、借款金额、借款用途、借款周期)。在删除了缺失值后,得到了7 599条完整记录的数据。在所选取的数据中,6 482条是没有违约的,1 117条是违约的。

第二,变量选取及处理。网贷违约风险预测中并没有标准的变量选取方法,本文通过参考国内外众多文献,最终确定了16个解释变量,可分为三大类,分别是借款人基本情况、借款产品信息以及借款人信用情况。1个预测变量,即是否违约。我们对各个变量进行了处理,具体(如表1所示)。

第三,实证结果。本文分别使用了支持向量机(SVM)、

表1　变量说明

	变量	变量说明
借款人基本信息	学历	高中及以下:0;大专:1;本科:2;研究生及以上:3
	性别	女:0;男:1
	是否有房	无:0;有:1
	是否有车	无:0;有:1
	是否有房贷	无:0;有:1

续表

	变量	变量说明
借款人基本信息	是否有车贷	无:0;有:1
	年龄	数值型变量
	婚姻状况	未婚:0;离异或寡居:1;已婚:2
	工作区域	西部:0;中部:1;东部:2
	工龄	小于 1 年:0;1—3 年:1;3—5 年:2;大于 5 年:3
	工资	0~5 000 元:0;5 000~10 000 元:1;10 000~20 000 元:2;20 000~50 000 元:3;大于 50 000 元:4
	就职企业规模	0—10 人:0;10—100 人:1;100—500 人:2 大于 500 人:3
借款产品信息	借款利率	数值变量
	借款周期	数值变量
	借款金额	数值变量
借款人信用情况	信用等级	HR=0;E=1;D=2;C=3;B=4;A=5;AA=6
预测变量	是否违约	没有违约=0;违约=1

决策树(DT)和随机森林(RF)进行预测,先在不同训练集下的建立模型,然后在测试集上进行预测,结果(如表 2 和表 3 所示)。

从以上数据可以得出,在进行违约率的预测时,三个数

表 2 **SVM**模型的在不同训练集下的预测正确率

训练集	70%	80%	90%
测试集总体预测正确率	87.93%	88.02%	88.83%
测试集未违约样本预测正确率	88.87%	89.09	89.78%
测试集违约样本预测正确率	82.93%	83.64%	83.72%

表 3 **DT**模型在不同训练集下的预测正取率

训练集	70%	80%	90%
测试集总体预测正确率	88.08%	89.01%	89.76%
测试集未违约样本预测正确率	88.76%	89.85%	89.90%
测试集违约样本预测正确率	83.39%	84.07%	84.22%

表 4　　RF模型在不同训练集下的预测正确率

训练集	70%	80%	90%
测试集总体预测正确率	90.04%	90.05%	90.02%
测试集未违约样本预测正确率	90.08%	90.12%	90.09%
测试集违约样本预测正确率	86.28%	87.86%	87.90%

据挖掘模型的预测效果都比较好，其中最好的是随机森林模型。同时，我们在训练集为90%的情况下，根据随机森林模型得出了各个变量的重要性程度。

各个解释变量的重要性依次为信用等级、借款数额、借款周期、借款利率、公司规模、工作时间、年龄、工资、学历、工作区域、婚姻状况、是否有车、是否有房、是否有房贷、性别、是否车贷。

四、结论与建议

第一，本文通过使用数据挖掘模型（支持向量机、决策树、随机森林）对网贷数据进行建模预测得出以下结论。首先，这三种模型对借款人的违约预测效果都很好，总体预测正确率都达到了87%以上，而对违约样本的预测正确率也都达到了82%以上，尤其以随机森林的预测效果最好，这可以为投资人和网贷平台在选择借款人时提供一些参考。其次，影响借款人违约的最重要的十个特征是借款人信用等级、借款数额、借款周期、借款利率、公司规模、工作时间、年龄、工资、学历、工作区域。

第二，基于以上结论，并结合中国P2P网贷行业发展现状，提出以下两点建议：首先，信用等级对借款人是否违约有重要的参考意义，所以网贷平台应该建立起一套标准的信用评级体系，能对借款人的信用等级进行有效的评分。其次，网贷平台间应该建立信息共享平台，使得平台间的征信信息能够得到有效共享，以降低违约风险。

参考文献：

[1] 刘丽丽.我国P2P网络借贷发展存在的风险及其监管对策[J].征信，2013，(11)：29-32.

[2] 李渊琦，陈芳.我国P2P网贷风险的风险分析及监管对策[J].上海金融，2015，(7)：78-81.

[3] 缪莲英，陈金龙.P2P网络借贷中社会资本对借款者违约风险的影响——以Prosper为例[J].金融论坛，2014，(3)：9-15.

[4] 刘鹏翔.P2P网贷平台借款人信用风险的影响因素分析——以拍拍贷平台为例[J].征信，2017，(3)：71-76.

[5] 李鑫.借款人声誉与风险识别——来自P2P网络借贷的证据[J].金融发展研究，2019，(6)：3-11.

[6] 程瑶.学历水平在借贷市场上的作用——来自P2P市场的经验证据[J].上海金融，2018，(2)：47-55.

[7] Hill R.C，W.E.Griffiths，G.C.Lim："Principles of econometrics"，Danvers，MA：John Wiley & Sons，Inc，2011.

[8] Goyal A.，R.Kau："Accuracy pre-diction for loan risk using machine learning models"，International Journal of Computer Science Trendsand Technology，2016，(1)：52-57.

Analysis on the Risk of Default of Network Loan Based on Data Mining

ZHONG Jiao-cong，FANG Hua

（University of Shanghai for Science and Technology，Shanghai 200093，China）

Abstract：Taking P2P network lending as an example，this paper selected 7559 records from personal loans from 2015 to 2018，analyzed borrowers default risk through data mining model，and identified the main factors affecting borrowers' default.These data mining models mainly include decision tree，support vector machine and random forest.The main conclusions include：firstly，using data mining model to predict default risk is very good，the best of which is random forest；secondly，the top five characteristics of importance are credit rating，loan amount，loan cycle，loan interest rate and working time of borrowers.

Key words：P2P network lending；data mining；default risk

[责任编辑　辰　敏]