DOI:10.15923/j.cnki.cn22-1382/t.2018.5.03

基于数据挖掘的小微商铺信用风险分析

程 晖, 董小刚*

(长春工业大学 数学与统计学院, 吉林 长春 130012)

摘 要: 通过 R 软件对原始数据进行预处理,剔除了原始数据缺失值,得到借款人的 16 个指标信息。对该数据建立 Logistic 回归、决策树、随机森林等数据挖掘模型,从准确性、正例命中率及可解释性等角度对比分析了上述模型,最终选取了逻辑回归模型作为小微商铺信用风险模型。逻辑回归模型下验证集的 ROC 值远大于 0.75,说明逻辑回归模型预测效果较好。

关键词: R 软件; P2P 信贷; 信用风险; Logistic 回归; 数据挖掘

中图分类号: O 212.1 **文献标志码:** A **文章编号:** 1674-1374(2018)05-0434-07

Analysis of credit risk of small and micro commercial shops based on data mining technology

CHENG Hui, DONG Xiaogang*

(School of Mathematics and Statistics, Changchun University of Technology, Changchun 130012, China)

Abstract: R program is used to preprocess the original data for eliminating missing values to obtain 16 indicators of 2816 borrower's information. data mining models such as Logistic regression, decision tree and random forest are established, and the models are analyzed considering the accuracy, hit rate and interpretability. The final Logistic regression model is determined as the micro-commercial shop credit risk model. ROC value of the Validation set under the Logistic regression model is much larger than 0.75, which indicates that the Logistic regression model is with rather accurate prediction.

Key words: R program; P2P credit; credit risk; Logistic regression; data mining.

0 引 言

近年来,互联网金融业发展迅猛。由于传统银行服务小微企业的收益与成本不大匹配,同时传统银行的业务办理速度较慢,以互联网为媒介的 P2P 信贷经营模式得到快速发展[1]。2005 年,英国尤努斯教授首次提出网络信贷服务平台概念[2]。P2P 信贷的产生,虽然给中小企业和个人带来了福音,实现了金融资源的优化配置,但是其自身也存在着巨大的风险和问题,由于监管上的空白,P2P 网贷各种携款潜逃、非法集资、高利贷等恶性事件时有发生。

收稿日期: 2018-08-25

基金项目: 国家自然科学基金资助项目(11301037, 11571151, 11671054); 吉林省教育厅"十三五"规划项目(2016316, 2016317)

作者简介: 程 晖(1994—),女,汉族,辽宁大连人,长春工业大学硕士研究生,主要从事高频数据方向研究,E-mail:1302005875@qq.com. *通讯作者:董小刚(1961—),男,汉族,吉林长春人,长春工业大学教授,博士生导师,主要从事数理统计方向研究,E-mail:dongxiaogang@ccut.edu.cn.

P2P 信贷平台存在巨大风险，信用风险研究角度也较为广泛。马运全[3]分别就网络借贷中逆向选择、道德风险和运作中存在的问题展开研究。艾金娣[4]分别就存在的制度风险和信用风险展开研究并提出相关防范措施。何晓玲[5]和付英军[6]认为主要是由于我国法律在这一块的空白带来的立法不完善和监管缺失的政策性风险，以及我国个人信用评价体系不健全带来的信息不对称下的信用风险和网络信贷平台自身建设的风险。虽然上述研究从不同角度分析了影响 P2P 网络信贷平台存在的风险因素，但是并没有更加细致地找出相关因素。文中使用 R 语言进行了全流程数据挖掘，选取的指标较为全面，并且使用了多种数据挖掘方法进行信用风险分析，可以普及到校园数据及各种大数据中；同时，文中旨在从借款人个人信用风险的角度去分析，为减少 P2P 网络信贷平台风险和完善我国 P2P 网络借贷行业治理提供有效建议。

1　数据来源及变量说明

近年来，信用风险的研究越来越多。姚凤阁[7]利用网络借贷平台中的借款人信息数据，选取了借款人信用等级、投标成功次数、投标流标次数、借款总额、利率、期限、每月还款、用户年龄、性别等 9 个指标，对 P2P 网络借贷平台借款人信用风险的影响进行分析，得出投标成功次数是影响借款人信用风险的最大因素且呈正相关，和借款人的借款期限、性别、年龄与借款人信用风险之间不存在相关关系的结论。方匡南[8]选取商业银行客户信用卡信贷数据中的家庭人口数、性别、年龄、婚姻、学历、职业、个人月收入、信用卡使用频率、客户是否为违约客户、信用卡张数、户籍所在地、所在地都市化程度、个人月开销占家庭月收入比例、月刷卡金额和家庭月收入共 15 个方面的信息来对 P2P 网络借贷平台借款人信用风险的影响进行分析。冯广庆[9]针对大学生群体选取了申请人的性别、年级、在校表现情况以及家庭状况来分析对大学生信用风险的影响，得出在大学生群体中女性比男性违约风险更高的结论。葛军[10]在信用卡信用风险研究中选取了申请人的性别、学历、年龄、婚姻、月收入、家庭人数、保险、职称、单位性质等 9 个指标变量，得出学历越高信用度越高和已婚者的违约概率比未婚者的违约概率低等结论。荣丽平[11]根据 P2P 网络借贷的特点，选取借款人年龄、性别、文化程度、工作年限、婚姻状况、月收入范围、是否购车、房产状况以及借款成功次数和逾期笔数等指标来对借款人的信用等级进行预测。

文中收集了文献[2]中附录的小微商铺信贷数据，通过剔除原始数据缺失值以及重复的样本数据信息，得到借款人的 16 个指标信息，对得到的数据通过数据挖掘技术进行分析，主要通过逻辑回归的方式进行分析，并且通过其他数据挖掘技术，如决策树、神经网络、随机森林、梯度提升等方式进行对比分析。这 16 个指标包括：是否为不良贷款、资产收益率、贷款原因/用途、信用记录中拖欠交易次数、店铺资产负债率比率、申请人学历、店铺经营时间、店铺年营业额、申请人信用记录、是否为本地户籍、申请人年龄、店铺月租金、申请人信用等级、店铺面积、雇员人数、所属行业。为了更好地度量 P2P 网络借贷平台的信用风险，文中用一个二值变量 Y(BAD)来表示因变量，即若为不良贷款，则 Y 用 1 表示；若非不良贷款，则 Y 用 0 表示。具体见表 1。

2　Logistic 模型介绍

在分析分类变量时，常常采用对数线性模型的方法，文中用的是对数线性模型中的 Logistic 模型。Logistic 模型[12]的优点在于它对自变量分布的假设条件没有限制，自变量可以是连续变量或离散变量；Logistic 模型中的因变量是一个二分类变量。

事件发生的概率为

$$p(y_i=1\mid x_i)=p[(\alpha+\beta x_i+\varepsilon_i)>C]=p[\varepsilon_i>(-\alpha-\beta x_i+C)]$$

当 $C=0$ 时，有

$$p(y_i=1\mid x_i)=p[\varepsilon_i\leqslant(\alpha+\beta x_i)]=\frac{1}{1+e^{-(\alpha+\beta x_i)}}$$

这个函数即为 Logistic 函数。

若将事件发生的概率 $p(y_i=1|x_i)$记为 p_i，则 p_i 表示第 i 个观测发生的概率，所以 Logistic 回归模型为

$$p_i=\frac{1}{1+e^{-(\alpha+\beta x_i)}}$$

则事件不发生的概率为

$$1-p_i=1-\frac{1}{1+e^{-(\alpha+\beta x_i)}}$$

所以，事件发生概率与不发生概率之比为

$$\frac{p_i}{1-p_i}=e^{(\alpha+\beta x_i)}$$

两边同时取对数即可将原先的非线性函数转换成一个线性函数，即

$$\ln\left(\frac{p_i}{1-p_i}\right)=\alpha+\beta x_i$$

表1 信用风险度量指标量化处理

符号	变量英文名	变量含义	变量说明
Y	BAD	是否为不良贷款	1:是,0:否
x_1	Profitrate	资产收益率	实际值(%)
x_2	Reason	贷款原因/用途	0:资金周转,1:扩大规模
x_3	Delinq	信用记录中拖欠交易次数	实际值
x_4	Debtinc	店铺资产负债率比率	百分比,全部负债除以全部资产
x_5	Education	申请人学历	1:高中及以下,2:大学,3:研究生及以上
x_6	Yropen	店铺经营时间	实际值(年)
x_7	Revenue	店铺年营业额	实际值(万元)
x_8	Creditage	申请人信用记录年份	实际值(年)
x_9	Local	是否为本地户籍	1:是,0:否
x_{10}	Age	申请人年龄	实际值
x_{11}	Rent	店铺月租金	实际值(万元)
x_{12}	Creditlevel	申请人信用等级	A^+:1,A:2,B^+:3,B:4
x_{13}	Storearea	店铺面积	实际值(m^2)
x_{14}	Numemployee	雇员人数	实际值
x_{15}	Indarea	所属行业	1:服务员,2:零售业,3:其他

在线性回归中，常采用最小二乘法和极大似然估计法估计未知总体参数，由于Logistic回归模型是一种非线性模型，最常采用的模型估计方法是极大似然估计法。

评价Logistic模型是否有效，通常是从两个方面来看，一方面是查看模型的拟合优度，即AIC准则和SBC准则，通常情况下，AIC和SBC取值越小，模型拟合得越好；另一方面是检查模型的预测准确性。

3 实证分析

通过数据分区的方式把原始数据分成训练集和验证集，比例为70%和30%，通过训练数据集训练模型，验证集来验证模型的效果。Logistic回归模型是文中使用的重要模型之一，Logistic回归模型[12-13]虽然对自变量分布的假设条件要求没那么高，但它对共线性却非常敏感，当自变量之间存在高度的自相关时，会导致估计的标准误差膨胀，故将应用Logistic回归模型时需对是否存在共线性进行检验。文中采用的是方差膨胀因子(VIF)作为是否存在多重共线性的判断标准，检验结果见表2。

表2 多重共线性检验

自变量	VIF
x_1	1.016
x_2	1.016
x_3	1.043
x_4	1.036
x_5	1.107
x_6	1.030
x_7	3.165
x_8	2.057
x_9	1.016
x_{10}	2.095
x_{11}	3.458
x_{12}	1.022
x_{13}	1.503
x_{14}	1.369
x_{15}	1.025

所有变量的逻辑回归结果见表 3。

表 3　所有变量的逻辑回归结果

回归系数	估计值	标准误差	z value	$Pr(>\|z\|)$
Intercept	−2.606	0.907	−2.873	0.004
x_1	−1.023	0.224	−4.567	0.000
x_2	0.152	0.224	0.681	0.496
x_3	0.447	0.078	5.733	0.000
x_4	0.080	0.132	6.060	0.000
x_5	−0.056	0.157	−0.358	0.720
x_6	−0.201	0.025	−7.988	0.000
x_7	−0.042	0.051	−0.824	0.410
x_8	−0.001	0.026	−0.024	0.980
x_9	0.131	0.223	0.584	0.559
x_{10}	−0.019	0.018	−1.067	0.286
x_{11}	−1.179	0.258	−4.573	0.000
x_{12}	0.637	0.106	6.037	0.000
x_{13}	0.011	0.006	1.726	0.084
x_{14}	0.105	0.097	1.081	0.279
x_{15}	0.096	0.124	0.772	0.440
AIC：622.91				

从检验结果可以看出，方差膨胀值（VIF）的平方根均小于 2，说明这 15 个自变量间不存在多重共线性问题。所以，可利用统计软件 R 将这 15 个自变量进行 Logistic 回归建模。从输出结果来看，有 2 816 条样本参与了建模。本次拟合出来的 Logistic 回归模型为：

$$\ln\left(\frac{p}{1-p}\right)=-2.605\ 6-1.022\ 9x_1+0.152\ 4x_2+0.447\ 3x_3+0.08x_4-0.056\ 2x_5-0.200\ 7x_6-0.042\ 3x_7-0.000\ 6x_8+0.130\ 6x_9-0.019x_{10}-1.179\ 5x_{11}+0.637\ 3x_{12}+0.010\ 9x_{13}+0.105\ 2x_{14}+0.095\ 9x_{15}$$

从参数的显著性检验结果可以得到，在这 15 个指标变量中只有 x_1，x_3，x_4，x_6，x_{11}，x_{12}，x_{13} 为显著非零，由于不显著变量较多，这里通过 AIC 准则进行变量选择，部分输出结果见表 4。

表 4　向后消除法回归汇总

回归系数	估计值	标准误差	z value	$Pr(>\|z\|)$
Intercept	−2.275	0.793	−2.870	0.004
x_1	−1.034	0.221	−4.683	0.000
x_3	0.433	0.078	5.699	0.000
x_4	0.081	0.013	6.121	0.000
x_6	−0.202	0.025	−8.114	0.000
x_{10}	−0.022	0.012	−1.783	0.075
x_{11}	−1.344	0.148	−9.079	0.000
x_{12}	0.636	0.105	6.082	0.000
x_{13}	0.015	0.005	2.897	0.000
AIC：612.23				

剔除 X_{10} 后的逻辑回归结果见表5。

表5 剔除 X_{10} 后的逻辑回归结果

回归系数	估计值	标准误差	z value	$Pr(>\|z\|)$
Intercept	−3.092	0.657	−4.707	0.000
x_1	−1.047	0.221	−4.739	0.000
x_3	0.438	0.078	5.609	0.000
x_4	0.082	0.013	6.198	0.000
x_6	−0.207	0.025	−8.266	0.000
x_{11}	−1.360	0.148	−9.216	0.000
x_{12}	0.649	0.104	6.227	0.000
x_{13}	0.015	0.005	2.851	0.000
AIC：613.45				

对于逐步回归的结果分析发现，x_{10} 不显著，去除 x_{10} 之后，模型的参数都显著，从而得到最终的模型，模型中包含了变量 x_1，x_3，x_4，x_6，x_{11}，x_{12}，x_{13}，所以最终的 Logistic 回归模型为：

$$\ln\left(\frac{p}{1-p}\right) = -3.091\,5 - 1.046\,6x_1 + 0.438\,4x_3 + 0.082x_4 - 0.206\,5x_6 - 1.360\,2x_{11} + 0.648\,7x_{12} + 0.014\,9x_{13}$$

从模型可以看出，对发生违约风险影响最大的是 x_{11}，其次是 x_1，再次是 x_{12}。通过上述参数估计可以计算出优比估计 Odds，见表6。

表6 优比估计

影响因素	Odds
x_1	0.351
x_3	1.550
x_4	1.085
x_6	0.813
x_{11}	0.257
x_{12}	1.913
x_{13}	1.015

由表6知：当 x_1 提高一个单位时，不良贷款的发生比为原来的0.351倍；当信用记录中 x_3 提高一个单位时，不良贷款的发生比为原来的1.550倍；当 x_4 提高一个单位时，不良贷款的发生比为原来的1.085倍；当 x_6 提高一个单位时，不良贷款的发生比为原来的0.813倍；当 x_{11} 提高一个单位时，不良贷款的发生比是原来的0.257倍；当 x_{12} 增加一个单位时，不良贷款的发生比将为原来的1.913倍(和表1信用等级B做比较)；当 x_{13} 提高一个单位时，不良贷款的发生比为原来的1.015倍。优比估计中点估计的值大于1，说明所选的自变量对事件的发生概率有正的作用。因此，x_3，x_4，x_{12}，x_{13} 对事件的发生概率有正的作用，即拖欠交易次数越多，店铺资产负债比越高(影响很小)，申请人信用等级越低，店铺面积越大(影响很小)，发生不良贷款的可能性越高；x_1，x_6，x_{11} 有负的作用，即资产收益率越高，店铺经营时间越长，店铺月租金越贵，发生不良贷款的可能性将会降低。

预测精度见表7。

表7 预测精度

Actual/predicted		predicted	
		norrmal	default
Actual	norrmal	778	2
	default	25	40
准确度：0.97			

从表7可以看出，预测精度为0.97，预测效果很好。

将验证数据集代入上述模型进行验证，得到

针对验证数据集的 ROC 曲线下面积 AUC 为 0.974。验证集的 ROC 曲线如图 1 所示。

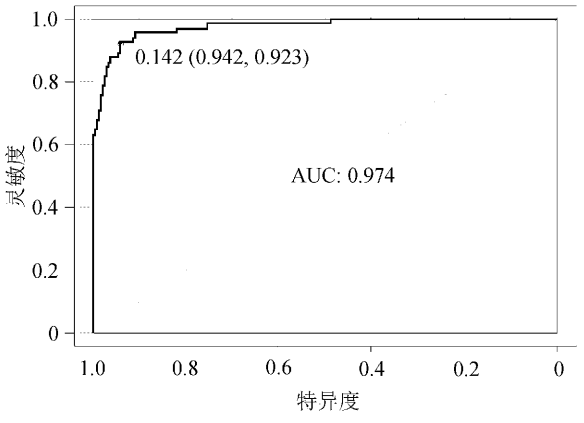

图 1　验证集的 ROC 曲线

一般认为 ROC 曲线下面积达到 0.75，该模型就具备了较好的预测能力。因此，从验证集上的 ROC 来看，模型拟合的预测效果非常好。

4　模型比较

除了使用逻辑回归进行建模，文中还采用决策树、随机森林、支持向量机等数据挖掘模型进行建模，通过对比模型的准确率、正例命中率、模型的可解释性及 ROC 曲线下的面积来进行模型选择。一般情况下，准确率、正例命中率和 ROC 曲线下面积介于 0～1 之间，取值越大越好，可解释性越强越好。因而综合各方面考虑，文中选取了逻辑回归作为最终的信用风险[14]模型。

模型比较见表 8。

表 8　模型比较

模型	准确率	正例命中率	可解释性	ROC 曲线下面积
逻辑回归	0.97	0.95	强	0.974
决策树	0.98	0.96	较强	0.965
随机森林	0.98	0.98	一般	0.995
神经网络	0.97	0.82	弱	0.941
支持向量机	0.98	0.98	弱	NA

5　结　语

通过采集 P2P 网络信贷平台上的借款人信息，选取了 2 816 条借款人是否不良贷款、资产收益率、贷款原因/用途、信用记录中拖欠交易次数、店铺资产负债率比率、申请人学历、店铺经营时间、店铺年营业额、申请人信用记录、是否为本地户籍、申请人年龄、店铺月租金、申请人信用等级、店铺面积、雇员人数、所属行业等 16 个指标信息，利用 R 软件进行 AIC 回归选择模型，最终得知资产收益率(x_1)、信用记录中拖欠交易次数(x_3)、店铺资产负债比率(x_4)、店铺经营时间(x_6)、店铺月租金(x_{11})、申请人信用等级(x_{12})、店铺面积(x_{13})这 7 个指标变量显著非零，再利用这 7 个变量进行 Logistic 回归建模，并对该模型的预测准确性进行检验，最后得出该模型的预测准确性为 0.97，并且模型验证集的 ROC 值远大于 0.75，预测效果较好。

所以，在 P2P 网络信贷平台上，出借人可以着重考虑小微商铺借贷人的资产收益率(x_1)、信用记录中拖欠交易次数(x_3)、店铺资产负债比率(x_4)、店铺经营时间(x_6)、店铺月租金(x_{11})、申请人信用等级(x_{12})、店铺面积(x_{13})这 7 个指标。一般来说，信用等级越低，不良事件发生的概率就越高；拖欠交易次数越多，发生不良贷款的可能性越高；店铺资产负债比越高，发生不良贷款的可能性相对较高。此外，资产收益率越高，发生不良贷款的可能性越低；店铺经营时间越长，发生不良贷款的可能性越低；店铺月租金越贵，发生不良贷款的可能性越低。这是由于店铺租金越贵(店铺一般处在繁华地段)，投入成本较多，需要大量流动资金周转。

参考文献：

[1]　李博.互联网金融的模型与发展[J].中国金融，2013(10)：19-21.

[2] 夏坤庄.深入解析SAS[M].北京:机械工业出版社,2015:586-615.
[3] 马运全.P2P网络借贷的发展、风险与行为矫正[J].新金融,2012(2):46-49.
[4] 艾金娣.P2P网络借贷平台风险防范[J].中国金融,2012,14:79-81.
[5] 何晓玲.P2P网络借贷现状及风险防范[J].中国商贸,2013,20:79-82.
[6] 付英军.P2P网贷行业发展存在的问题及对策[J].当代经济,2015,22:28-29.
[7] 姚凤阁.P2P网络借贷平台借款人信用风险影响因素研究:来自"拍拍贷"的经验依据[J].哈尔滨商业大学学报,2016(1):3-10.
[8] 方匡南.基于Lasso-logistic模型的个人信用风险预警方法[J].数量经济技术经济研究,2014(2):125-136.
[9] 冯广庆.基于Logistic模型的大学生信用卡风险研究[J].知识经济,2011,14:55.
[10] 葛君.基于Logistic模型的信用卡信用风险研究[J].中国信用卡,2010,24:26-32.
[11] 荣丽平.P2P网络借贷个人信用风险评估[J].财会月刊,2015,35:94-96.
[12] 王小宁.R语言实战[M].北京:人民邮电出版社,2015:280-299.
[13] 于立勇.基于Logistic回归的违约概率预测研究[J].财经研究,2004(9):15-23.
[14] 刘克,冯松.金融市场风险识别与控制[J].长春工业大学学报:社会科学版,2005,17(1):12-14.