

# P2P 网贷个人信用评价指标体系的构建

石澄贤, 陈雪交

**摘要:** 为增加 P2P 网贷平台信用评价的可信性, 改善因信息不对称而导致的企业信用风险问题, 结合实际业务需求从表征信息、行为信息及状态信息三个方面选择个人信用评价指标变量。通过计算变量的 WOE (Weight of Evidence)、IV (Information Value) 值初步观测变量对目标的显著程度, 再用 SAS 软件对整体变量进行逻辑回归。以 IV 和逻辑回归相结合的方式筛选指标, 遴选出 22 个指标作为 P2P 网贷平台信用评价体系指标。这种个人信用评估指标的遴选结果保留了信息量大, 对信用评价贡献概率大的指标。

**关键词:** P2P 网贷平台; 个人信用评价; 指标体系; 逻辑回归

**作者简介:** 石澄贤, 常州大学数理学院教授, 硕士生导师; 陈雪交, 常州大学数理学院硕士研究生。

**中图分类号:** F832.479

**文献标识码:** A

**Doi:** 10.3969/j.issn.2095-042X.2016.01.012

P2P 网络借贷是互联网技术与民间借贷相结合的一种金融创新模式, 是出借人通过网络借贷平台将资金贷给借款人的一种民间借款方式, 属于无抵押贷款<sup>[1]</sup>。由于借贷双方信息不对称, 使得 P2P 网贷过程中的个人信用评价成为影响借贷交易的重要因素。然而中国缺少高度透明的个人信用体系, 借款人的信用评价完全由 P2P 网络借贷平台的信用认证机制做出<sup>[2]</sup>。而信用认证机制中最重要的一个环节就是评价指标的选取, 如果评价指标过于单一, 那么评价模型对其风险揭示作用就会存在局限性。为了进一步发挥信用认证机制的风险揭示作用并降低借贷双方信息不对称问题, 应丰富和完善多层次认证指标体系并构建合理的信用指标。

关于个人信用评估指标体系, 我国已有许多学者进行了深入研究。就信用评估指标考察的内容而言, 主要涉及还款能力与还款意愿两个方面, 当然也有学者在此基础上提出了一些新的观点, 如郭昱等<sup>[3]156</sup>在还款意愿和还款能力的基础上增加了考察成长潜力的信用评价指标体系; 赖辉等<sup>[4]</sup>提出了“个人信用行为状态”概念, 从信用行为和信用状态以及行为状态的关联上构建个人信用评估指标; 蒋小兔等<sup>[5]</sup>主要从小额信贷的实际开展情况和市场特征上选取评价指标; 就评估方法而言, 主要涉及逻辑回归、决策树及支持向量机等, 如张国政等<sup>[6]</sup>通过逻辑回归测得影响个人信用风险的关键因素是借款人的年龄、婚姻情况、受教育程度等六项指标; 孙同阳等<sup>[7]</sup>通过决策树模型对信用风险进行评价; 夏晗<sup>[8]</sup>通过主成分分析和支持向量机相结合的方法对信用风险进行预测。

纵览众多文献发现, 虽然目前评估指标在内容上开始偏向还款意愿及行为状态方面, 但整体考察的力度并不是那么大, 在信用指标的选取上多数还是更强调还款能力而很少涉及还款意愿及个人的行为状态信息。然而当具有还款能力时, 借款人是否违约就取决于其还款意愿的强弱<sup>[9-10]</sup>及个人的行为状态。因此对还款意愿及行为状态的考察是极其重要的。此外, 在评估的方法上, 更多文献采用单一的评估方法, 由于评估方法本身的局限性可能导致选出的指标存在一些不合理性问题。因此结合以上分析, 本文在评

估内容上从表征信息、行为信息及状态信息三个方面构建个人信用评价指标体系，使还款能力与还款意愿相结合，行为信息与状态信息相结合，更加全面细致地考察信用评估指标；在评估方法上采用 IV 与逻辑回归相结合的方法选取指标，从而使选出的指标体系无论是在内容上还是在方法上都更加合理可信。

一、个人信用评价指标设计

个人信用评价指标体系的构建不仅要全面客观，还要充分考虑所选指标是否符合 P2P 网贷的特点<sup>[3]157</sup>。通过借鉴当今 P2P 网贷平台信用评价指标体系的研究结果和考虑 P2P 网贷平台自身的特点及实际业务需求，从表征信息、行为信息以及状态信息三个方面，构建了涵盖定性与定量指标相结合的 P2P 网贷平台个人信用评价指标体系。

本文中的表征信息主要涉及最基本的性别、年龄等信息，这些指标对客户的行为预测并不具有因果关系，但是根据历史数据统计可得到一定的规律。行为信息主要涉及产品类型、申请频次、申请金额等，行为是内部需求在外部特定环境下的一种表现，是内部需求的结果，对客户的行为预测可以表现出相关性。而状态信息是指客户的社会经济状态和社会网络关系，对预测客户行为具有一定的因果关系<sup>[11]</sup>。

本文数据来源于融 360 金融数据分析大赛，根据以上指标设计原则，结合实际业务需求，首先从 135 个变量中初步筛选了 39 个评价指标构成该指标体系，如表 1 所示；然后通过计算 WOE，IV 值及对变量进行逻辑回归显著性检验后对指标进一步筛选。

表 1 初选信用指标体系

	指标	描述	IV	IV_RANK	缺失值占比
表征信息	op_type	职业类型	0.094	8	0.00%
	user_work_period	工作年限	0.078	10	0.00%
	user_age	用户年龄	0.018	17	0.00%
	company_type	公司类型	0.060	11	0.00%
	qid123	文化程度	0.019	15	0.00%
	qid137	现居住地居住时间	0.003	36	0.00%
	qid133	性别	0.005	33	0.00%
	qid140	行业	0.009	21	0.00%
	user_nationality	户籍	0.005	31	0.00%
	qid122	婚姻状况	0.007	25	0.00%
行为信息	qid139	居住类型	0.010	20	0.00%
	product_type	产品类型	0.143	6	3.07%
	pv	总 pv	0.008	22	57.20%
	pv_index_loan	贷款首页 PV	0.005	32	57.20%
	pv_apply_total	申请次数	0.006	30	57.20%
	pv_ask	问答页 PV	0.005	34	57.20%
	pv_calculator	计算器页 PV	0.006	29	57.20%
	order_count_loan	贷款下单量	0.016	19	57.20%
	pv_daikuan	贷款总 PV	0.008	23	57.20%
	pv_credit	信用卡总 PV	0.007	26	57.20%
	pv_search_daikuan	搜索页 PV	0.006	28	57.20%
	pv_detail_daikuan	贷款详情页 PV	0.006	27	57.20%
	date	日期	0.019	16	57.20%
	term	申请期限	0.057	12	0.00%
	limit	申请金额	0.086	9	0.00%
	guarantee_type	产品担保类型	0.208	1	0.00%
	loan_quota_max	最大额度	0.194	2	0.00%
	loan_quota_min	最小额度	0.165	3	0.00%
	repayment_type	还款方式	0.155	5	0.00%

续表 1

	指标	描述	IV	IV_RANK	缺失值占比
状态信息	user_loan_experience	两年内征信情况	0.022	14	0.00%
	user_has_car	是否有车	0.007	24	0.00%
	qid77	是否有公积金	0.098	7	0.00%
	cash_receipts	现金收入	0.003	35	0.00%
	user_income_by_card	打卡收入	0.156	4	0.00%
	col_value	房产价值	0.003	37	0.00%
	com_month_flow	月流水	0.016	18	0.00%
	car_value	车产价值	0.002	39	0.00%
	qid79	信用卡最高额度	0.002	38	0.00%
	spam_score	欺诈评分	0.033	13	3.07%

## 二、个人信用指标选取的理论依据

本文主要通过 IV 与逻辑回归相结合的方法确定指标显著性,再结合实际业务需求选取最终指标。涉及到的关键理论知识如下所示:

### (一) WOE 计算方法

由于样本数据中连续变量较多,为降低变量属性的个数,并且平滑变量的变化趋势,通常会先对变量的取值进行分箱并计算每个变量属性的 WOE (Weight of Evidence) 值<sup>[12]</sup>。WOE 主要是通过比较不同变量属性之间的差异度来反映自变量对因变量的影响程度,原理是如果同一变量的不同属性得出的 WOE 差异度越大,那么说明不同的变量取值对目标变量的区分度越大,进而说明该变量对目标变量的作用越显著。计算公式如下:

$$woe_i = \ln\left[\left(\frac{g_i}{g}\right) / \left(\frac{b_i}{b}\right)\right] \quad (1)$$

式(1)中的  $woe_i$  是某变量第  $i$  个属性对应的 WOE 值;  $g_i$  是某变量第  $i$  个属性对应的好客户数;  $b_i$  是某变量第  $i$  个属性对应的坏客户数;  $g$  是样本中的总的好客户数;  $b$  是样本中总的坏客户数。

### (二) IV 计算方法

如果说 WOE 是自变量取某一个值时对目标变量的影响,那么 IV (Information Value) 则衡量的是某一个变量的信息量,从公式来看的话,相当于是自变量 WOE 值的一个加权求和,其值的大小决定了自变量对于目标变量的影响程度;这种影响程度可以从另一个角度来理解,式(2)中的  $g_i/g$  与  $b_i/b$  可以理解为某个自变量关于目标变量的条件密度,当这两个条件密度距离越远时,说明这个自变量对目标变量的辨识度越好。因此,只要 IV 值大就说明对目标变量影响显著。计算公式如下:

$$IV = \sum_i \left(\frac{g_i}{g} - \frac{b_i}{b}\right) \times \ln\left[\left(\frac{g_i}{g}\right) / \left(\frac{b_i}{b}\right)\right] = \sum_i \left(\frac{g_i}{g} - \frac{b_i}{b}\right) \times woe_i \quad (2)$$

式(2)中的  $woe_i$  是某变量第  $i$  个属性对应的 WOE 值;  $g_i$  是某变量第  $i$  个属性对应的好客户数;  $b_i$  是某变量第  $i$  个属性对应的坏客户数;  $g$  是样本中的总的好客户数;  $b$  是样本中总的坏客户数。

### (三) 逻辑回归模型及变量筛选

逻辑回归模型主要针对目标变量是分类变量构建的回归模型,其数学模型如下:

假设在自变量  $x_1, x_2, \dots, x_n$  作用下,以  $Y=1$  表示某事件发生,概率为  $p$ ,  $Y=0$  表示该事件不发生,概率为  $(1-p)$ ,  $p/(1-p)$  为发生概率和不发生概率之比,记做“优势”(odds),若对 odds 取自然对数,得到:  $\text{logit}(Y) = \ln(\text{odds}) = \ln(p/1-p)$  称为  $Y$  的 logit 变换<sup>[13]233-234</sup>,则 logistic 回归模型为:

$$\text{logit}(Y) = \ln\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n \quad (3)$$

式 (3) 中,  $x_1, x_2, \cdots, x_n$  是自变量,  $b_1, b_2, \cdots, b_n$  是自变量对应的回归系数,  $b_0$  为常数项。

以  $x_1, x_2, \cdots, x_{35}$  分别表示表 1 中 35 个指标, 以  $Y=1$  表示个人信用好,  $Y=0$  表示个人信用坏, 利用式 (3) 可以构建个人信用好坏的 logistic 回归模型式。对于 logistic 回归模型式 (3) 是否要包含或删除某预测变量  $x_i$ , 即为变量筛选问题。目前主要的变量筛选方法有向前回归法、向后回归法、逐步回归法和全模型回归法。本文主要采用逐步回归法进行变量筛选。将变量逐个引入模型, 每引入一个解释变量后都要进行 F 检验, 并对已经选入的解释变量逐个进行 t 检验。当原来引入的解释变量由于后面解释变量的引入变得不再显著时, 则将其删除, 以确保每次引入新的变量之前回归方程中只包含先主动变量<sup>[13]238</sup>。

三、遴选 P2P 网贷个人信用评价指标

首先根据公式 (1)、(2) 将所有数据用 SAS 软件计算其对应的 IV 值 (如表 1 所示), 并用逐步进入的方法对所有变量进行逻辑回归, 得出最终的变量显著性结果如表 2 所示:

表 2 逻辑回归最终确定的变量

变量	描述	卡方	p 值	变量	描述	卡方	p 值
cash_receipts	现金收入	55.95	<0.01	term	申请期限	199.37	<0.01
user_income_by_card	打卡收入	65.16	<0.01	qid123	文化程度	58.75	<0.01
user_work_period	工作年限	6.9	<0.01	qid133	性别	57.34	<0.01
user_age	用户年龄	35.79	<0.01	qid140	行业	89.94	<0.01
com_month_flow	月流水	17.74	<0.01	product_type	产品类型	285.86	<0.01
spam_score	欺诈评分	133.94	<0.01	user_loan_experience	两年内征信情况	145.99	<0.01
order_count_loan	贷款下单量	189.72	<0.01	user_has_car	是否有车	33.77	<0.01
pv_credit	信用卡总 PV	9.1	<0.01	qid77	是否有公积金	216.05	<0.01
date	日期	220.29	<0.01	repayment_type	还款方式	19.88	<0.01
op_type	职业类型	68.32	<0.01				

通过对比表 1 和表 2 结果可知, 除 cash\_receipts (现金收入)、pv\_credit (信用卡总点击量)、qid133 (性别) 和 user\_has\_car (用户是否有车) 这四个变量在 IV 表中的排名稍微靠后一些外, 其余变量基本上在 IV 中排序都比较靠前, 说明 IV 结果和逻辑回归得出的结果具有一致性, 而根据实际业务理解, “现金收入” “信用卡总点击量” 和 “用户是否有车” 对是否批贷是有显著影响的。此外参考其他学者大量实验结果发现, “性别” 对是否批贷也是有显著影响的。逻辑回归中加入这 4 个变量是很符合实际的, 另外, limit (申请金额)、qid122 (婚姻状况)、qid139 (居住类型) 在 IV 中的排名比较靠前, 而在逻辑回归中并没有把其加入模型中。根据实际业务分析, “申请金额” 对是否批贷具有显著作用, “婚姻状况” 和 “居住类型” 通常也会对是否批贷产生一定影响, 因此我们最终从 39 个变量中选出以下 22 个变量作为个人信用评价指标体系的最终指标, 结果如表 3 所示。

表 3 个人信用评价最终指标体系

	指标	描述	指标	描述
表征信息	op_type	职业类型	qid133	性别
	user_work_period	工作年限	qid122	婚姻状况
	user_age	用户年龄	qid139	居住类型
	qid123	文化程度	qid140	行业

续表 3

	指标	描述	指标	描述
行为信息	product _ type	产品类型	term	申请期限
	order _ count _ loan	贷款下单量	repayment _ type	还款方式
	pv _ credit	信用卡总 PV	limit	申请金额
	date	日期		
状态信息	user _ loan _ experience	两年内征信情况	user _ income _ by _ card	打卡收入
	user _ has _ car	是否有车	com _ month _ flow	月流水
	qid77	是否有公积金	spam _ score	欺诈评分
	cash _ receipts	现金收入		

#### 四、结论

目前 P2P 行业处于快速发展阶段, 借款需求相对旺盛, 但对信用风险防控的要求也逐渐提高。如何构建合理可信的信用风险评估模型, 选取全面客观的信用指标体系, 从而降低个人信用风险并减少 P2P 行业损失, 成为摆在人们面前的一道难题。此外, 由于我国的金融监管机构并未出台对 P2P 的监管细则, P2P 行业仍处于无准入门槛、无行业标准、无主管机构的三无状态<sup>[14]</sup>。加之, 个人信用机制的不健全以及一些借款人为了获得借款提供虚假的信用材料, 无形中增加了 P2P 行业信用风险评估及评价指标选取的难度。为了在某种程度上解决这一难题, 也为了能给更多工作者在选取指标时提供一定参考, 本文在借鉴诸多学者研究成果的基础上进一步提出了 P2P 信用评价指标选取的方法。

从文中表 1 的 IV 值结果来看, 对目标变量影响程度比较大的几个指标基本上是行为信息或状态信息中的指标。以往人们对个人信用指标的研究主要集中在表征信息这类基本信息的考察, 对行为信息和状态信息的考察并不太重视, 但由以上结果可以看出, 行为信息和状态信息对个人信用评价有着至关重要的作用, 有些指标的显著程度比表征信息中的指标还要显著。因此, 本文建议在对个人信用指标进行筛选时加大对行为信息和状态信息的考察。最后, 本文 P2P 网贷平台信用评价体系指标遴选以信息量大, logistic 回归模型分析对信用评价贡献概率大为原则。这样选取的个人信用评估指标保留了重要和核心指标。可以保持指标选取的客观全面, 确保评估方法的合理可信性。这种通过对指标的实际数据进行分析, 尽量做到不重不漏、科学合理的方法为评估关键指标的选取提供了有用的标准。这种在许多指标中遴选重要指标的办法希望能够给实际工作者提供有益的参考。

#### 参考文献:

- [1] Mingfeng Lin, N R Prabhala, Siva Viswanathan. Judging borrowers by the company they keep: social networks and adverse selection in online Peer-to-Peer lending [J]. Journal of Women's Health, 2009
- [2] 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, 28 (4): 137.
- [3] 郭昱, 马翻翻, 郑超文. 我国小微企业信用评价指标体系的构建 [J]. 金融经济, 2015, 22 (2).
- [4] 赖辉, 帅理, 周宗放. 个人信贷客户信用评估的一种新方法 [J]. 技术经济, 2014, 33 (9): 97-103.
- [5] 蒋小兔, 查奇芬. 常州市小额信贷信用风险评价研究 [J]. 中国集体经济, 2014, 30 (16): 84-85.
- [6] 张国政, 陈维煌, 刘呈辉. 基于 Logistic 模型的商业银行个人消费信贷风险评估研究 [J]. 金融理论与实践, 2015, 34 (3): 53-57.
- [7] 孙同阳, 谢朝阳. 基于决策树的 P2P 网贷信用风险评价 [J]. 商业经济研究, 2015, 34 (2): 81.
- [8] 夏晗. 基于主成分分析和支持向量回归机组合模型的电子商务信用风险度预测研究 [J]. 现代情报, 2015, 35 (1): 76-79.
- [9] 彭红枫, 叶永刚. 基于资本监管要求和还款意愿的贷款定价研究 [J]. 中国管理科学, 2009, 17 (2): 8-14.
- [10] 彭红枫, 叶永刚. 基于还款能力和还款意愿的贷款定价研究 [J]. 中国管理科学, 2011, 19 (6): 41-47.



- [11] 常国珍. 胸有成竹! 数据分析的 SASEG 进阶 [M]. 北京: 电子工业出版社, 2015: 136-137.
- [12] 杨池然. SAS 开发经典案例解析 [M]. 北京: 机械工业出版社, 2013: 315.
- [13] 姚志勇. SAS 编程与数据挖掘商业案例 [M]. 北京: 机械工业出版社, 2013.
- [14] 潘庄晨, 邢博. 我国 P2P 网络借贷模式的发展现状及风险揭示研究 [J]. 未来与发展, 2014, 38 (6): 86-89.

## The Construction of P2P Network Lending Personal Credit Evaluation Index System

Shi Chengxian, Chen Xuejiao

**Abstract:** In order to increase the credibility of P2P network lending platform and solve the problem of enterprise credit risks caused by information asymmetry, personal credit evaluation index variables are selected in three aspects, namely, representation information, behavior information and status information combined with actual business needs. By calculating WOE and IV of variables, the significance of variables to targets in the preliminary observation is showed. The logistic regression of all variables is presented by use of SAS software. Combing IV and logistic regression, 22 variables are selected as P2P network lending platform credit evaluation indexes. Through the selection, indexes which contain a large amount of information and make a lot of contribution to credit evaluation are reserved.

**Key words:** P2P network lending platform; personal credit evaluation; index system; logistic regression

(收稿日期: 2015-10-21; 责任编辑: 沈秀)