基于*K-S*检验与距离相关分析的网络借贷信用评价指标体系构建

段 翀

（内蒙古财经大学 经济学院，呼和浩特010070）

摘 要：网络借贷作为一种新型互联网金融模式，提升了金融资源使用效率，缓解了小企业融资难的困局。构建合理的网络借贷信用评价指标体系，从而对网络借贷的潜在风险及时甄别与预防，对互联网金融健康持续发展意义重大。本文根据*K-S*检验与距离相关分析相结合，筛选对借款客户违约状态甄别能力强的指标，建立了网络借贷信用评价指标体系，通过P2P网络借贷（peer to peer lending，个人对个人借贷）平台LendingClub交易数据进行实证研究，结果表明：不仅借款金额、借款利率等借款标的特征对借贷者违约具有显著相关性，借款者年龄等个人特征、借款者年收入等财务特征以及借款者违约次数等信用特征均对借贷者违约风险产生显著影响。投资者在出借资金时，往往青睐于已婚、年龄适中、具有一定工作经历、历史违约次数较少的借款人。因此，风险监管部门应构建网络借贷违约风险评估模型，对P2P平台进行风险监测，同时建立关键信息共享机制，融合多源数据，明确审查范围，实现P2P网络借贷行业健康有序发展。

关键词：网络借贷；信用评价；违约显著区分；*K-S*检验；距离相关系数

中图分类号：F830.5 **文献标志码**：A **文章编号**：1002—980X（2020）5—0035—14

一、研究背景

P2P网络借贷（peer to peer lending，个人对个人借贷）系指资金出借方与借入方不是通过银行而是依托互联网平台建立借贷关系的无抵押贷款[1]。

网络借贷作为一种极具活力的新型互联网金融模式，具有无需中介、交易便利与覆盖面广等优点，备受借贷双方青睐，近年来获得了蓬勃发展。根据网贷之家发布的《2018年中国网络借贷行业年报》显示，截至2018年中国网络借贷行业累计平台达6430家，全年累计交易规模为17948亿元，行业参与人数突破千万。

网络借贷突破了时间与地点的局限，提升了金融资源的使用效率，缓解了小企业融资难的困局。然而，由于贷款门槛较低、缺乏专业信贷人员与借贷双方缺乏现实接触等因素，使得相比于传统信贷，网络借贷的信息不对称情况更严重，导致平台违约事件频发、信用风险日益加剧。信用风险过大已成为网络借贷发展的瓶颈。

科学评估网络借贷的信用风险，从而对网络借贷这一新经济业态的潜在风险隐患及时甄别与预防，对于互联网金融健康持续发展意义重大。网络借贷信用风险评估已成为备受关注的前沿领域[2]。

评价指标体系建立是网络借贷信用评价的基础环节，若是构建的评价指标体系包含大量重复或是不重要的指标，则运用何种评价模型均无法得到科学的评价结果。

目前网络借贷的信用评价主要侧重两类。

（一）信用评价指标体系研究现状

Francesco等[3]通过相关分析筛选指标，建立了包括盈利能力、偿付能力、流动性状况和信贷质量共4个准则的网络借贷信用评级指标体系。谭中明等[4]通过网络问卷调查方式与隶属度分析结合，利用因子分析方法，从流动性、透明度、品牌等方面构建网贷风险评价指标体系。蒋翠清等[5]运用信息增益、信息增益率及卡方检验对软、硬信息指标的重要性进行排序，通过封装筛选确定最优的网络信贷指标体系。张成虎与武博华[6]在网络调研与因子分析筛选基础上，构建了包含软信息的P2P网络借贷信用风险评价指标体系。刘传

收稿日期：2019—12—24

基金项目：国家自然科学基金“基于违约金字塔原理的小企业贷款定价模型”（71171031）；国家自然科学基金青年科学基金“基于不均衡支持向量机的小企业信用风险评价理论与模型”（71201018）；内蒙古自然科学基金面上项目“基于基准利率风险补偿的小企业贷款定价模型”（2017MS0709）

作者简介：段翀（1981—），男，内蒙古呼和浩特人，博士，内蒙古财经大学讲师，研究方向：信用评价。

哲等[7]以对称不确定性为基础测算指标间的相关系数，删除冗余指标，并利用异质集成模型对网贷信用评分问题进行研究。

现有信用评价指标体系的共同不足：一是现有信用评价指标筛选方法基本采用相关分析、因子分析剔除反映信息冗余的指标，上述方法仅仅揭示了变量间的线性关联程度，但P2P网络借贷作为一种互联网金融创新模式，其海量、复杂的借贷数据往往具备非线性特征。因此，采用现有相关分析、因子分析无法揭示网贷信用评价指标间的非线性联系，从而导致信用评价指标筛选的结果不可靠；二是现有信用评价指标不能保证对客户违约与否进行有效甄别，这与信用风险评价这一根本目的相悖。信用评价目的是甄别违约企业与非违约企业。

（二）网络借贷影响因素的研究现状

Puro等[8]通过美国网络借贷平台prosper数据实证发现，网络借贷成功率与借贷利率及借贷额正相关。Lin等[9]发现借款者的朋友关系可以降低借款者承担的利率并降低违约风险。廖理等[10]发现虽然借贷利率越高，借款者才容易获得贷款，但此类贷款的违约风险也更高。Emekter等[11]认为信用评分、借款期限、负债收入比等对于借款违约概率有重要影响。Malekipirbazari和Aksakalli[12]认为借款者的债务收人比也是影响借贷行为的一个关键因素。Lin和Viswanathan[13]发现文化与地理位置是影响借贷重要因素，贷款者更倾向于借贷给地域接近、文化相近的借款者。何光辉等[14]运用Logistic与Probit模型对中国网络借贷风险决定因素进行分析。李杰和刘露[15]根据Logistic模型发现借款者总收入、总支出是网贷违约与否关键因素。李延喜等[16]运用Logistic与Cox模型发现，借贷成功并不完全取决于贷款利率，借款者的年龄、学历及婚姻状况均有重要影响。

现有研究方法的不足之处：现有网络借贷影响因素中不仅涉及诸多不能显著甄别客户违约状态的指标，还涵盖不少信息重复的指标。

上述问题，本文在网络借贷信用评价海选指标体系，采用*K-S*检验选取可以显著甄别借款人违约状态的指标，进而通过距离相关分析剔除掉反映信息重复的指标，最终构建网络借贷的信用评价指标体系。并通过全球最大的P2P网络借贷平台LendingClub的实际交易数据进行实证研究。

二、网络借贷信用评价指标体系构建原理

（一）P2P网络借贷的特点

（1）借款者和投资者之间不存在真实的接触，信息更加不透明，导致投资者对借款者的信用风险进行评价更加困难。

（2）现有关于银行各类贷款表现的研究[17-18]表明，消费者信用贷款在很大程度上会受到国内生产总值、失业率和利率等宏观因素的影响，因此如何控制这些宏观因素对于P2P网络借贷信用风险的影响，将是一个重要的问题。

（3）P2P网络借贷作为依托云计算、社交网络等新兴技术平台涌现出来的互联网金融创新模式，海量、复杂的网络借贷数据往往具备非线性、高维的大数据特征。

（二）网络借贷信用评价的问题

问题1：怎样从众多繁杂指标中遴选得到可以对网络借贷者违约状态显著分辨的重要指标。

问题2：如何克服现有指标筛选方法仅仅反映指标间线性关联程度的弊端。现有信用评价指标筛选方法基本采用相关分析、因子分析剔除反映信息冗余的指标，上述方法仅揭示了变量间线性关联程度，但P2P网络借贷依托于云计算、社交网络等新兴技术平台，其数据量极大，往往具备非线性特征。

（三）解决问题的思路

问题1的解决思路：将企业数据分为违约和非违约两类样本，通过*K-S*检验比较违约样本与非违约样本的分布函数是否有显著差异，按照*K-S*检验统计值越大、违约样本分布函数和非违约样本的分布函数的偏离愈大，指标越能显著甄别客户的违约状态，选取可以显著区分违约状态与否的指标。

问题2的解决思路：距离相关分析是近年来高维数据非线性相关分析的流行度量方法，其从特征函数的距离角度定义了两个随机变量间的非线性相关系数。本文采用距离相关系数反映指标间的线性与非线性的综合关联程度，在关联程度强的一对指标中，剔除*K-S*检验较小、对违约状态影响较小的指标，删除了反映信

息冗余指标。

（四）评价指标筛选原理

通过K-S检验统计值的大小反映指标对违约状态分辨能力上的差异，按照K-S检验统计值越大、违约样本与非违约样本的分布函数的偏离越大，指标分辨客户违约与否的能力就越强。进而根据距离相关分析在两个关联程度高的指标中筛选出鉴别分辨违约状态能力强的指标。克服现有相关分析、因子分析等指标筛选方法仅揭示了指标间的线性关联程度和无法反映指标间的非线性关联程度的弊端，弥补现有研究不以能否区分违约状态为标准遴选评价指标的不足。

网络借贷信用评价指标体系构建原理如图1所示。

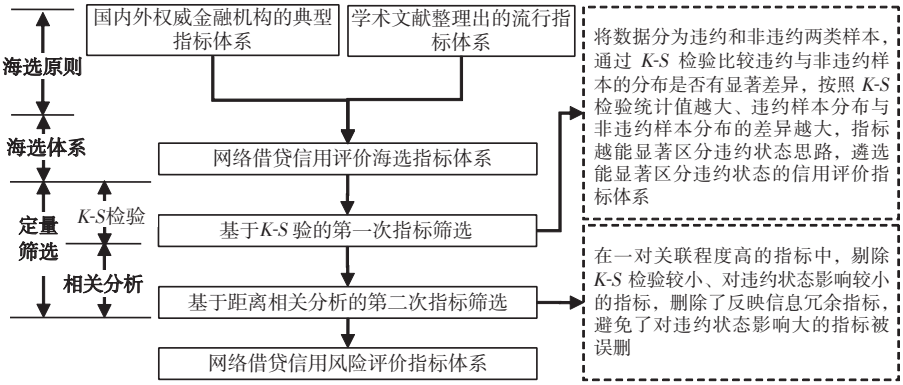

图1　网络借贷信用评价指标体系构建原理

三、网络借贷信用评价指标体系的构建

（一）海选指标体系的建立

1. 网络借贷信用风险的内涵

网络借贷信用风险指网络借贷中贷款者未能按照合同约定及时足额还本付息而给资金出借方带来的风险损失[13-15]。

资金出借方通过考量贷款人的还款能力及还款意愿这两方面来评估其发生违约的可能性。贷款人的还款能力可以通过贷款者年收入等财务特征与贷款者职业等个人特征来反映；而贷款人的还款意愿可由贷款者违约次数等个人信用特征来体现。

同时，现有研究[8-16]表明：贷款金额、贷款利率等借款特征也对网络贷款的信用风险影响显著。此外，现有关于银行各类贷款表现的研究[17-18]表明，消费信用贷款在很大程度上会受到国内生产总值、失业率和利率等宏观因素的影响，因此将外部宏观经济特征纳入网络借贷信用评价指标体系中。

2. 准则层设置

一是通过设置借款者个人特征、借款者财务特征两个准则反映借款者还款能力；二是通过设置个人信用特征准则反映借款者还款意愿；三是通过设置宏观经济特征、借款特征两个准则反映外部经济环境、借款状况对借款者清偿能力的影响。

3. 海选指标体系的构建

以网络借贷信用风险内涵为基础，根据国内外网络借贷信用评价文献的流行高频指标[3-16]，建立了包括借款金额、年龄等指标构成的涉及借款标的特征、借款者个人特征、借款者财务特征、借款者信用特征及宏观经济特征5个准则层的网络借贷信用评价的海选指标体系，见表1。

（二）信用评价指标体系的建立

1. 指标数据的归一化

指标数据归一化是把指标原始数据转化为[0，1]间的数，剔除单位及量纲对评价结果的影响。

信用评价指标可分成定量指标与定性指标。定量指标分为成本类型指标、效益类型指标及区间型指标。

成本型指标系指网络借款者的信用状况与指标的数值负相关，即指标数值愈大，则说明借款者的信用状况愈差。效益型系指标系指网络借款者的信用状况与指标的数值正相关，即指标数值愈大，则说明借款者的

表 1 网络借贷信用评价海选指标体系

序号	准则层	指标名称	指标类型	文献来源	筛选方式
1	借款标的特征	X_1借款金额	成本型	[3,5-6,8]	**保留**
2		X_2借款利率	成本型	[4-6,8]	**保留**
3		X_3还款月数	成本型	[1,4-6]	显著性删除
4		X_4还款月支付额	成本型	[3-6,8]	显著性删除
5	借款者个人特征	X_5借款者年龄	区间型	[2-6,8]	**保留**
6		X_6借款者性别	定性	[2-4]	显著性删除
7		X_7借款者学历	定性	[3-6]	距离相关删除
8		X_8借款者职业	定性	[5-7]	**保留**
9		X_9借款者居住地	定性	[6,8]	显著性删除
10		X_{10}借款者婚姻状况	定性	[1,6,8]	**保留**
11	借款者财务特征	X_{11}年收入	效益型	[1-3,5]	**保留**
12		X_{12}认证年收入	效益型	[4-6,8]	距离相关删除
13		X_{13}月工资收入	效益型	[1,4-6]	显著性删除
14		X_{14}债务收入比	成本型	[3-6,8]	**保留**
15		X_{15}年总支出	成本型	[1-4]	显著性删除
16		X_{16}是否有房贷	定性	[3,8]	距离相关删除
17		X_{17}是否有车贷	定性	[2-5,7]	显著性删除
18		X_{18}工作年限	定性	[3-6,8]	**保留**
19		X_{19}是否拥有自有住房	定性	[2-6,8]	距离相关删除
20		X_{20}拥有不动产价值总额	效益型	[2-4]	显著性删除
21	借款者信用特征	X_{21}信用评级	定性	[1,6,8]	**保留**
22		X_{22}两年内违约次数	定性	[1-3,5]	**保留**
23		X_{23}信用年龄	定性	[1-4]	显著性删除
24		X_{24}收入是否经过认证	定性	[3,8]	显著性删除
25		X_{25}信用总额度	效益型	[2-5,7]	距离相关删除
26		X_{26}信用使用率	效益型	[2-4]	距离相关删除
27		X_{27}信用账户总数	定性	[3-6]	显著性删除
28	宏观经济特征	X_{28}股市收益率	效益型	[5-7]	**保留**
29		X_{29}居民消费价格指数	区间型	[4-6]	距离相关删除
30		X_{30}人均GDP增速	效益型	[1,7-8]	显著性删除
31		X_{31}人均可支配收入	效益型	[3-6,8]	显著性删除
32		X_{32}失业率	成本型	[2-6,8]	**保留**

信用状况愈好。

成本类、效益类指标归一化公式[12]如下所示。

$$x_{ij}=\frac{\max\limits_{1\leqslant j\leqslant n}p_{ij}-p_{ij}}{\max\limits_{1\leqslant j\leqslant n}p_{ij}-\min\limits_{1\leqslant j\leqslant n}p_{ij}} \tag{1}$$

$$x_{ij}=\frac{p_{ij}-\min\limits_{1\leqslant j\leqslant n}p_{ij}}{\max\limits_{1\leqslant j\leqslant n}p_{ij}-\min\limits_{1\leqslant j\leqslant n}p_{ij}} \tag{2}$$

其中：x_{ij}为第i个指标第j个借款的归一化值；p_{ij}为第i个指标第j笔借款原始数据；n为借款个数。

区间型指标是指当指标的数据值落在某一个特定区间内都是合理的指标。例如：居民消费价格指数、年龄等两指标。指标“居民消费价格指数”理想区间是[100.6, 104.7][17-18]。“居民消费价格指数”数值处于该区间中既不通货膨胀又不通货紧缩。根据对网络借贷平台发放调查问卷，发现将指标“年龄”合理区间范围设置为[30,48]，即年龄处于该区间的借款者还款意愿、清偿能力都是最强的。

设q_1为指标最佳区间左端点，q_2为指标最佳区间右端点，根据区间指标的归一化打分公式[12]如下：

$$x_{ij}=\begin{cases}1-\dfrac{q_1-p_{ij}}{\max(q_1-\min\limits_{1\leqslant i\leqslant n}v_{ij},\max\limits_{1\leqslant i\leqslant n}v_{ij}-q_2)}, & p_{ij}<q_1\\ 1-\dfrac{p_{ij}-q_2}{\max(q_1-\min\limits_{1\leqslant i\leqslant n}p_{ij},\max\limits_{1\leqslant i\leqslant n}p_{ij}-q_2)}, & p_{ij}>q_2\\ 1, & q_1\leqslant p_{ij}\leqslant q_2\end{cases} \tag{3}$$

定性指标的标准化得分是在对网贷信用评价专家进行实地访谈调研基础上，按照定性指标的不同程度确定量化打分标准。见表2。

表2　定性指标打分标准

指标	打分标准	打分分值
学历	具有本科及以上学历	1.0
	具有高中以上且本科以下学历	0.5
	其他	0.0
…	…	…
借款者2年内违约次数	违约0次	1.0
	违约1次	0.5
	违约2次及以上	0.0
借款者工作年限	工作年限≥10年	1.0
	5年≤工作年限<10年	0.7
	2年≤工作年限<10年	0.5
	0年≤工作年限<2年	0.2

2. 指标的正态分布检验

通过Jarque-Bera正态检验，可判断指标是否服从正态分布。Jarque-Bera正态检验步骤[19]如下。

(1)建立假设检验。

原假设：第i个指标X_i服从正态分布(H0)。

备择假设：第i个指标X_i不服从正态分布(H1)。

(2)构造JB统计量(即Jarque-Bera检验统计量)。设$\bar{x}_i$为第i个指标标准化得分的平均值，n为样本借款总数，x_{ij}为第i个指标第j个借款的标准化得分，j=1,2,…,n，则

$$\bar{x}_i=\frac{1}{n}\sum_{j=1}^{n}x_{ij} \tag{4}$$

设$\bar{\sigma}_i$为第i个指标的标准化得分的标准差，则

$$\bar{\sigma}_i=\frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)^2 \tag{5}$$

设S_i为第i个指标偏度系数，由文献[19]可知

$$S_i=\frac{1}{n}\sum_{j=1}^{n}\left(\frac{x_{ij}-\bar{x}_i}{\bar{\sigma}_i}\right)^3 \tag{6}$$

式(6)中其他字母含义与式(4)、式(5)相同。

设K_i第i个指标的峰度系数，由文献[19]可知

$$K_i=\frac{1}{n}\sum_{j=1}^{n}\left(\frac{x_{ij}-\bar{x}_i}{\bar{\sigma}_i}\right)^4 \tag{7}$$

设JB_i为第i个指标Jarque-Bera检验统计量，则

$$JB_i=\frac{n}{6}\left[S_i^{\,2}+\frac{(K_i-3)^2}{4}\right] \tag{8}$$

式(8)中其他字母含义与式(6)、式(7)相同。

(3)检验标准[19]。原假设H0成立时，Jarque-Bera检验统计量JB_i服从自由度为2的χ^2分布[19]，给定显著性水平α，查表可得χ^2分布的临界值J_0。若统计量JB_i大于临界值J_0，则拒绝原假设H0，即第i个指标X_i不服从正态分布；反之，则接受原假设H0，即第i个指标X_i服从正态分布。

3. 违约显著区分的指标筛选方法

通过该方法可删除对违约状态区分不显著的指标。

按照某个指标数据，把借款数据分为违约和非违约两类，若该指标K-S检验值愈大，即违约的经验分布与非违约的经验分布的偏离愈大，说明评价指标甄别借款人是否违约的能力越强，指标对信用评价结果影响显著，应保留；反之，说明该指标无法有效区分违约借款人与非违约借款人，指标对信用评价结果影响不大，须删除。

K-S检验筛选指标的计算步骤如下。

(1)建立假设检验。

原假设:第i个指标的违约样本的分布与非违约样本的分布没有显著差异(H0)。

备择假设:第i个指标的违约样本分布与非违约样本分布有显著差异(H1)。

(2)构造K-S检验的D统计量。

步骤1:两类样本经验分布函数的确定。

以违约样本的经验分布为例。设违约借款个数为n_1,非违约借款个数为n_2,借款总数为n,$n=n_1+n_2$。

令$x_{i1},x_{i2},\cdots,x_{i,n1}$为第$i$个指标$n_1$个违约借款的标准化值。将这$n_1$个数从小到大排序,重新编号得到这$n_1$个标准化值的次序统计量$x^1_{i,(1)}\leqslant x^1_{i,(2)}\leqslant\cdots\leqslant x^1_{i,(n_1)}$。同理得到$n_2$个非违约借款的标准化值的次序统计量$x^2_{i,(1)}\leqslant x^2_{i,(2)}\leqslant\cdots\leqslant x^2_{i,(n_2)}$。

设$F_i^{(1)}(x)$为第i个指标在n_1个违约借款中的经验分布函数,k为在第i个指标n_1个违约借款的标准化值中小于等于x的标准化值的个数,$x^1_{i,(j)}$为第i个指标在n_1个违约借款中的第j个次序统计量,$j=1,2,\cdots,n_1$,由文献[19]可知

$$F_i^{(1)}(x)=\begin{cases}0, & x<x^1_{i,(1)}\\ \dfrac{k}{n_1}, & x^1_{i,(k)}\leqslant x<x^1_{i,(k+1)}\\ 1, & x\geqslant x^1_{i,(n_1)}\end{cases}\tag{9}$$

同理得第i个指标的非违约经验分布$F_i^{(2)}(x)$。

$$F_i^{(2)}(x)=\begin{cases}0, & x<x^2_{i,(1)}\\ \dfrac{v}{n_2}, & x^2_{i,(v)}\leqslant x<x^2_{i,(v+1)}\\ 1, & x\geqslant x^2_{i,(n_2)}\end{cases}\tag{10}$$

其中:v表示在第i个指标n_2个非违约借款的标准化值中,小于等于x的标准化值的个数。

步骤2:K-S检验D统计量的确定。

设D_i为第i个指标K-S检验的D统计量,I为由第i个指标n个借款的标准化值构成的实数集合,即$I=\{x_{i1},x_{i2},\cdots,x_{i,n}\}$,$F_i^{(1)}(x)$为第$i$个指标的违约经验分布,$F_i^{(2)}(x)$为第$i$个指标的非违约经验分布,由文献[19]可知

$$D_i=\max_{x\in I}\left|F_i^{(1)}(x)-F_i^{(2)}(x)\right|\tag{11}$$

其中:第i个指标K-S检验统计量D_i等于第i个指标违约经验分布$F_i^{(1)}(x)$与第i个指标非违约经验分布$F_i^{(2)}(x)$之差的绝对值的最大值。

K-S检验统计量D_i反映了第i个指标区分违约状态的能力强弱。第i个指标的D统计量越大,第i个指标在违约样本与非违约样本中的差异越大,表明第i个指标区分违约状态能力越强;反之亦然。

式(11)采用K-S检验筛选指标的好处:一是按照违约样本与非违约样本的分布函数的差异越大,这个指标越能显著区分违约与否状态的思路,构造指标的K-S检验值,遴选能显著区分违约状态的指标,弥补现有研究不以能否区分违约状态为标准遴选评价指标的不足;二是采用K-S检验这一对评价指标的总体分布无任何要求、适用于分布未知的非参数统计方法筛选指标,克服现有方法要求指标服从正态分布的这一严格假设弊端。

(3)筛选标准。在原假设H0成立时,第i个指标K-S检验的统计量D_i服从Kolmogorov分布[17]。给定显著性水平α,通过查表可得Kolmogorov分布的临界值D_0。

①若统计量D_i大于等于临界值D_0,则拒绝原假设H0,即第i个指标的违约样本分布与非违约样本分布有显著差异,说明违约样本与非违约样本能被第i个指标明显区分,则保留第i个指标。

②若统计量D_i小于临界值D_0,则接受原假设H0,即第i个指标的违约样本分布与非违约样本分布没有显著差异,说明违约样本与非违约样本不能被第i个指标明显区分,则删除第i个指标。

4. 冗余信息剔除的指标筛选方法

该方法可在关联程度高的一对指标中筛选出违约甄别能力强的指标,确保得到信息不重复的指标。

距离相关系数是一种新型相关系数,其基本思想是根据两个随机变量的联合分布函数$F(x,y)$与各自的

边缘分布函数$F_X(x)$、$F_Y(y)$间的距离测度随机变量X与Y之间的相关性[20-21]。与传统皮尔逊相关系数、秩相关系数等线性相关系数相比，距离相关系数无论变量间是线性关系或是非线性关系均可度量，无需任何假设与分布条件，具有很强的普适性。因此，本文采用距离相关系数度量同一准则层下两指标间的相关性，进而进行冗余指标的删除。

距离相关系数筛选指标的步骤如下。

(1)距离相关系数的计算。设有m个指标，n个借据。令$\boldsymbol{X}_i$为第i个指标归一化值的向量，即$\boldsymbol{X}_i=\left(x_{i1}, x_{i2}, \cdots, x_{in}\right)$。则向量$\boldsymbol{X}_i$与向量$\boldsymbol{X}_j$的距离相关系数$dr_{ij}$[20]为

$$dr_{ij}=\frac{d\mathrm{cov}\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)}{\sqrt{d\mathrm{cov}\left(\boldsymbol{X}_i, \boldsymbol{X}_i\right)\times d\mathrm{cov}\left(\boldsymbol{X}_j, \boldsymbol{X}_j\right)}} \tag{12}$$

其中：$d\mathrm{cov}\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)$为向量$\boldsymbol{X}_i$、$\boldsymbol{X}_j$的距离协方差，由下式(13)计算得到。

$$d\mathrm{cov}\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)=\frac{1}{n^2}\sum_{k,l=1}^{n}\left(A_{kl}B_{kl}\right) \tag{13}$$

其中：A_{kl}、B_{kl}由式(14)与式(15)确定。

$$A_{kl}=a_{kl}-\bar{a}_{k\cdot}-\bar{a}_{\cdot l}+\bar{a}_{\cdot\cdot} \tag{14}$$

$$B_{kl}=b_{kl}-\bar{b}_{k\cdot}-\bar{b}_{\cdot l}+\bar{b}_{\cdot\cdot} \tag{15}$$

设x_{ik}为第i个指标第k笔借款的归一化值，则式(14)的4个参数分别由式(16)~式(19)确定。

$$a_{kl}=\left|x_{ik}-x_{il}\right| \tag{16}$$

$$\bar{a}_{k\cdot}=\frac{1}{n}\sum_{l=1}^{n}a_{kl} \tag{17}$$

$$\bar{a}_{\cdot l}=\frac{1}{n}\sum_{k=1}^{n}a_{kl} \tag{18}$$

$$\bar{a}_{\cdot\cdot}=\frac{1}{n}\sum_{k,l=1}^{n}a_{kl} \tag{19}$$

(2)临界值确定。设定临界值$M\in[0,1]$，若距离相关系数绝对值$\left|r_{ik}\right|\geq M$，则删除两个指标中反映违约鉴别能力弱的指标。两个指标距离相关系数大于0.8时属高度相关[20]。因此，选取临界值$M=0.8$。

(3)指标筛选标准。若第i个指标与第k个指标的距离相关系数绝对值$\left|r_{ik}\right|\geq 0.8$时，则第$i$个指标与第$k$个指标反映信息重复，应删除其中K-S检验值较小的；反之，说明指标反映信息不重复，同时保留两指标。

本研究筛选信息冗余指标方法的好处：通过距离相关系数删除反映信息重复的指标，保留K-S检验值大的，即对违约区分能力强的指标，避免对违约状态区分能力强的指标被误删，无论指标间是线性关系或是非线性关系均可适用，无须指标数据满足正态分布，适用于指标分布未知的情形。

(三)评价指标体系的合理性检验

信用评价指标体系合理与否是看基于指标体系构建信用评价模型的违约预测力是否显著。即检验利用指标体系构建的信用评价模型违约预测能力越显著，信用评价指标体系就越合理。

先利用上述指标筛选方法构建网络借贷的信用评价指标体系。再根据该指标体系及Logistic模型可以计算得到每个借款者的违约概率(PD_i)。将PD_i与违约临界值比较，可对借款者是否违约进行预测。继而采用ROC曲线(受试者工作特征曲线)AUC值(ROC曲线所覆盖的区域面积)对网贷信用评价指标体系的违约预测效果进行检验。

将实际违约借款被模型判定为违约借款数量记为DD；实际违约借款被模型判定为非违约的借款数量记为DN；实际非违约借款被模型判定为违约的数量记为ND；实际非违约借款被模型判定为非违约的数量记为NN，见表3。

表3　实际违约状态与模型判别结果划分

实际违约状态	模型判别结果		
	违约(1)	非违约(0)	合计数
违约(1)	实际违约借款判定为违约DD	实际违约借款被判定为非违约DN	实际违约借款总数
非违约(0)	实际非违约借款判定为违约ND	实际非违约借款判定为非违约NN	实际非违约借款总数
合计	判别结果为违约的借款总数$DD+ND$	判别结果为非违约的借款总数$DN+NN$	全部网络借款总数

ROC曲线涉及两个变量，灵敏度（*Sensitivity*）和特异度（*Specificity*），如式（20）和式（21）[22]所示：

$$Sensitivity=DD/(DD+DN) \tag{20}$$

灵敏度（*Sensitivity*）等于实际违约借款中被模型判定为违约的个数*DD*与实际违约借款总数（*DD+DN*）的比率，即借款违约状态的判对率。

$$Specificity=ND/(ND+NN) \tag{21}$$

特异度（*Specificity*）等于实际非违约的借款中被模型判定为非违约的个数*ND*与实际非违约借款总数（*ND+NN*）的比率，即借款非违约状态判对率。

ROC曲线的纵轴为*Sensitivity*，横轴即1−*Specificity*，也就是1−借款非违约状态判对率。

ROC曲线下方围成面积为AUC值。当横轴不变时，纵轴越向上，即实际违约借款判对率越高，模型判别准确率也越高，ROC曲线也越向上，曲线下围成的面积AUC值也越大。因此，AUC值越大，信用评价模型对违约状态判别准确性越高，信用评价指标体系也就越合理。

四、实证研究

（一）实证样本与数据来源

本文的指标实证样本来自美国P2P借贷平台LendingClub提供的借款标的数据[23]，样本数据区间为2009—2014。Lending Club成立于2007年，是目前世界上最大的在线P2P网络借贷平台，平台提供企业及个人信贷、房贷及消费贷款等借贷品种。在样本区间内，选取已完结的网络借款，并去掉数据缺失较多的指标，最终得到31000条借款信息，对应31000个借款人。非违约样本为27000个，违约样本为4000个。违约系指贷款到期后90天内未能足额偿还贷款的本金与利息。P2P网络借贷的指标原始数据见表4第31000列所示。表4第33行为贷款借款的违约状态标识，违约、非违约分别用“1”和“0”标识。

（二）信用评价指标体系的建立

1. 指标的归一化

（1）定量指标归一化。根据表4的指标类型，分别将表4第1～31000列的正向指标、负向指标、区间型指标数据p_{ij}代入式（1）～式（3），得到指标的标准化得分x_{ij}，列入表4后31000列各定量指标对应行。

（2）定性指标标准化。根据表4的指标类型及表2定性指标的打分标准，为表4中的各个定性指标进行归一化打分。结果列入表4各定性指标的对应行。

2. 指标的正态分布检验

（1）Jarque-Bera正态检验统计值的确定。以第1个指标X_1“借款金额”的Jarque-Bera正态检验统计量的确定过程为例。把表4第1行指标X_1的归一化得分x_{1j}、借款总数n=31000依次代入式（4）～式（7），得到指标X_1的偏度系数S_1=−0.531、峰度系数K_1=−0.949。把偏度系数S_1=−0.531、峰度系数K_1=−0.949代入式（8），得到指标X_1的正态检验统计值JB_1=126.412。将结果列入表5第1行第3列。同理得其余指标的统计量JB_i，结果列入表5第3列其余行。

（2）正态检验结果。原假设H0成立时，第i个指标的检验统计量JB_i服从自由度为2的χ^2分布[18]，给定显著性水平α=0.05，查表得χ^2分布的临界值J_0=5.991。由于表5第3列的81个指标的JB_i均大于J_0=5.991，由正态检验标准，则32个指标X_i均不服从正态分布。在表5第4列用“否”标注。

由于所有32个指标X_i均不服从正态分布，故本文采用*K-S*检验、距离相关分析的非参数统计方法筛选信用评价指标。

3. 违约显著区分的指标第1次筛选

（1）*K-S*检验值的确定。以指标“X_1借款金额”为例。

步骤1：违约样本经验分布函数的确定。把表4第1行指标X_1的4000个违约借款的归一化值x_{1j}按照从小到大次序排列，得到指标X_1对应次序统计值$x^1_{1,(1)}=0\leqslant x^1_{1,(2)}\leqslant\cdots\leqslant x^1_{1,(4000)}=0.914$。把得到的指标$X_1$标准化值对应的次序统计值$x^1_{1,(1)}$、…、$x^1_{1,(32)}$，$n_1$=4000代入式（9），得到违约样本经验分布$F_1^{(1)}(x)$。仿照上述过程，可得非违约样本经验分布$F_1^{(2)}(x)$。

步骤2：*K-S*检验值的确定。把指标X_1第1个借款的归一化值x_{11}依次代入函数$F_1^{(1)}(x)$、$F_1^{(2)}(x)$，得到x_{11}的违约样本经验分布值$F_1^{(1)}(x_{11})$、非违约样本经验分布值$F_1^{(2)}(x_{11})$，得到$\left|F_1^{(1)}\left(x_{11}\right)-F_1^{(2)}\left(x_{11}\right)\right|$。同理，可得

表4　P2P网络借贷指标筛选原始数据

序号	准则层	指标	指标类型	网贷客户的指标原始数据p_{ij}			网贷客户的指标数据归一化结果x_{ij}					
							违约借款客户			非违约借款客户		
				（1）客户1	…	（31000）客户31000	（1）客户1	…	（4000）客户4000	（4001）客户4001	…	（31000）客户31000
1	C_1借款标的特征	X_1借款金额（万元）	成本型	11.800	…	10.500	0.486	…	0.304	0.177	…	0.517
2		X_2借款利率（%）	成本型	9.680	…	10.070	0.626	…	0.411	0.201	…	0.380
3		X_3还款月数	成本型	12	…	6	0.750	…	0.400	1	…	0.850
4		X_4还款月支付额	成本型	0.983	…	1.750	0.422	…	0.300	0.175	…	0.203
5	C_2借款者个人特征	X_5借款者年龄	区间型	34	…	29	1.000	…	0.600	0.850	…	0.900
6		X_6借款者性别	定性	男	…	男	1.000		0.500	1.000	…	1.000
7		X_7借款者学历	定性	本科	…	高中	1.000		0.500	1.000	…	0.500
8		X_8借款者职业	定性	事业单位人员	…	服务业从业人员	1.000		0.5000	0.700	…	0.700
9		X_9借款者居住地	定性	本地居住地	…	本地居住地	1.000	…	0.000	0.500	…	1.000
10		X_{10}借款者婚姻状况	定性	已婚	…	已婚	1.000	…	1.000	0.500	…	1.000
11	C_3借款者财务特征	X_{11}年收入（万元）	效益型	10.000	…	8.900	0.700	…	0.639	0.826	…	0.639
12		X_{12}认证年收入（万元）	效益型	9.000	…	8.000	0.572	…	0.438	0.724	…	0.478
13		X_{13}月工资收入（万元）	效益型	0.600	…	0.550	0.673	…	0.697	0.941	…	0.537
14		X_{14}债务收入比	成本型	9.070	…	6.946	0.317	…	0.217	0.142	…	0.581
15		X_{15}年总支出（万元）	成本型	7.470	…	7.031	0.468	…	0.810	0.692	…	0.571
16		X_{16}是否有房贷	定性	有	…	有	0.500	…	0.500	1.000	…	0.500
17		X_{17}是否有车贷	定性	无	…	无	1.000	…	1.000	0.500	…	1.000
18		X_{18}工作年限（年）	定性	8	…	10	0.700		1.000	0.500	…	1.000
19		X_{19}是否拥有自有住房	定性	是	…	是	1.000	…	0.500	0.500	…	1.000
20		X_{20}拥有不动产价值总额（万元）	效益型	52.382	…	57.500	0.644	…	0.503	0.495	…	0.705
21	C_4借款者信用特征	X_{21}信用评级	定性	2	…	8	0.900	…	0.100	0.600	…	1.000
22		X_{22}两年内违约次数	定性	0	…	0	1.000	…	1.000	0.500	…	1.000
23		X_{23}信用年龄	定性	6	…	4	0.500	…	1.000	0.700	…	0.500
24		X_{24}收入是否经过认证	定性	是	…	是	1.000	…	1.000	1.000	…	1.000
25		X_{25}信用总额度	效益型	23150	…	14870	0.650	…	0.787	0.026	…	0.529
26		X_{26}信用使用率	效益型	26.400	…	52.750	0.743	…	0.031	0.865	…	0.775
27		X_{27}信用账户总数	定性	8	…	12	0.509	…	0.549	0.507	…	0.518
28	C_5宏观经济特征	X_{28}股市收益率	效益型	12.9	…	13.5	0.575	…	0.648	0.615	…	0.262
29		X_{29}居民消费价格指数	区间型	101.400	…	105.4	1.000	…	0.651	1.000	…	0.931
30		X_{30}人均GDP增速	效益型	0.009	…	0.009	0.143	…	0.635	1.000	…	0.143
31		X_{31}人均可支配收入增速	效益型	−0.0006	…	−0.0006	0.042	…	0.424	0.693	…	0.042
32		X_{32}失业率	成本型	6.800	…	7.600	0.633	…	0.716	0.559	…	0.455
33	—	是否违约y_i	—	0.000	…	0.000	1	…	1	0	…	0

表5　Jarque-Bera正态检验结果

序号	准则层	指标	JB_i	结果
1	C_1借款标的特征	X_1借款金额（万元）	126.412	否
…		…	…	…
4		X_4还款月支付额（万元）	12.66	否
5	C_2借款者个人特征	X_5借款者年龄	1411.89	否
…		…	…	…
10		X_{10}借款者婚姻状况	358.96	否
11	C_3借款者财务特征	X_{11}年收入（万元）	23.31	否
…		…	…	…
20		X_{20}拥有不动产价值总额（万元）	555.23	否
21	C_4借款者信用特征	X_{21}信用评级	16.13	否
…		…	…	…
27		X_{27}信用账户总数	902.03	否
28	C_5宏观经济特征	X_{28}股市收益率	764.63	否
…		…	…	…
32		X_{32}失业率	894.26	否

其余归一化值的违约样本与非违约样本经验分布值之差的绝对值。

综上，总共得到指标 X_1 的 31000 个、违约样本与非违约样本的经验分布函数值之差的绝对值 $\left|F_1^{(1)}(x_{11})-F_1^{(2)}(x_{11})\right|,\cdots,\left|F_1^{(1)}(x_{1,31000})-F_1^{(2)}(x_{1,31000})\right|$。求解这 31000 个绝对值中的最大值即得到指标 X_1 的 *K-S* 检验统计值 D_1，即 $D_1=\max_{x\in I}\left|F_1^{(1)}(x)-F_1^{(2)}(x)\right|=1.473$。其中，$I$ 为由表 4 第 1 行指标 X_1 的归一化值组成的实数集合。结果列入表 6 第 1 行第 3 列。重复上述步骤 1～步骤 2，可得其余 31 个指标的 *K-S* 检验统计值 D_i，结果列入表 6 第 3 列其余行。

（2）*K-S* 检验筛选指标的结果。*K-S* 检验统计量 D_i 服从 Kolmogorov 分布[19]。给定显著性水平 $\alpha=0.05$，通过查表可得 Kolmogorov 分布的临界值 $D_0\geqslant 1.358$。

通过表 6 第 3 列可知，在 32 个指标中，“X_3 还款月数”等 13 个指标的 *K-S* 检验值全都低于 1.358，检验不通过，说明这些指标的违约借款与非违约借款的经验分布函数并不存在明显区别，指标无法显著甄别借款者是否违约，应删除。通过表 6 第 3 列可知，32 个指标中，“X_1 借款金额”等 19 个指标的 *K-S* 检验值全都大于 1.358，表明这些指标的违约借款与非违约借款的经验分布函数存在明显区别，指标可以显著甄别借款者是否违约，应保留。

4. 冗余信息删除的指标第 2 次筛选

（1）距离相关系数的确定。经过第四节第（二）节第 3 小节的第一次筛选，删除了 13 个指标，剩余 19 个指标。将表 6 第 1～2 列的 *K-S* 检验保留的 19 个指标按准则层合并，合并后的 19 个指标进行距离相关分析的第二次指标遴选。用于距离相关分析的第二次指标遴选的 19 个指标标准化数据见表 7。利用表 7 的标准化数据以及式（12）～式（19），可得到同一准则层下两个指标的距离相关系数。

以指标 X_1 的和 X_2 的距离相关系数计算为例。利用指标 X_1 归一化得分 x_{1j}、指标 X_2 归一化得分 x_{2j} 及式（12）～式（19），得到指标 X_1 和 X_2 的距离相关系数 $r_{12}=0.353$。其他指标的距离相关系数类推可得。将所有指标中绝对值大于 0.8 的距离相关系数值 r_{ik} 列入表 8 第 5 列。表 8 第 2、4 列的 D_i 来源于表 6 第 3 列的相应行。

（2）距离相关分析筛选结果。选取 0.8 作为距离相关系数临界值。由表 8 第 5 列可知，共有 7 对指标的距离相关系数大于 0.8，故此 7 对指标属于反映信息冗余，在这 7 对指标中保留 *K-S* 检验统计值 D_i 较大的指标。由表 6 第 3 列的 D_i 可知，$\max\{D_{11}、D_{12}\}=\{2.311、1.842\}=2.311$，指标 X_{11}“年收入”的 *K-S* 检验统计值 D_{11} 最大，故保留指标 X_{11}，删除指标 X_{12}，删除的指标列于表 8 第 2 行第 6 列。同理，其他的删除指标列于表 8 第 6 列的其他行。

综上，通过距离相关分析将 *K-S* 检验筛选后留下的 19 个指标进行第二次筛选，去除反映信息冗余的 7 个指标，最终保留了 12 个指标。删除的 7 个在表 1 用“冗余信息删除”标出；最终保留的 12 个指标在表 1 用“保留”标出。

表 6　*K-S* 检验指标筛选结果

序号	(1)准则层	(2)指标	(3)D_i值	(4)结果
1	C_1借款标的特征	X_1借款金额（万元）	1.473	保留
2		X_2借款利率（%）	1.524	保留
3		X_3还款月数	0.749	删除
4		X_4还款月支付额（万元）	0.604	**删除**
5	C_2借款者个人特征	X_5借款者年龄	1.894	保留
6		X_6借款者性别	0.690	**删除**
7		X_7借款者学历	1.597	保留
8		X_8借款者职业	2.032	保留
9		X_9借款者居住地	0.307	**删除**
10		X_{10}借款者婚姻状况	3.161	保留
11	C_3借款者财务特征	X_{11}年收入（万元）	2.311	保留
12		X_{12}认证年收入（万元）	1.842	保留
13		X_{13}月工资收入（万元）	0.722	**删除**
14		X_{14}债务收入比	2.017	保留
15		X_{15}年总支出（万元）	0.814	**删除**
16		X_{16}是否有房贷	1.610	保留
17		X_{17}是否有车贷	0.502	**删除**
18		X_{18}工作年限（年）	1.884	保留
19		X_{19}是否拥有自有住房	1.716	保留
20		X_{20}拥有不动产价值总额（万元）	0.873	**删除**
21	C_4借款者信用特征	X_{21}信用评级	2.301	保留
22		X_{22}两年内违约次数	2.174	保留
23		X_{23}信用年龄	0.218	**删除**
24		X_{24}收入是否经过认证	0.733	**删除**
25		X_{25}信用总额度	1.690	保留
26		X_{26}信用使用率	1.879	保留
27		X_{27}信用账户总数	0.437	**删除**
28	C_5宏观经济特征	X_{28}股市收益率	2.631	保留
29		X_{29}居民消费价格指数	1.780	保留
30		X_{30}人均GDP增速	0.596	删除
31		X_{31}人均可支配收入增速	0.627	**删除**
32		X_{32}失业率	1.461	保留

5. 网络借贷信用评价指标体系的建立

在32个指标中，根据*K-S*检验去掉区分违约不显著的13个指标，利用距离相关分析删除7个信息冗余的指标，最后建立了包含12个指标的小企业信用评价指标体系，见表9第$a\sim d$列。同时，最终保留12个指标在表1中用“保留”标出。

表7　进行距离相关分析的19个海选指标的标准化数据

序号	准则层	指标	31000笔网络借款的海选指标归一化数据v_{ij}					
			4000笔违约借款			27000笔违约借款		
			(1)借款1	…	(4000)借款4000	(4001)借款4001	…	(31000)借款31000
1	C_1借款标的特征	X_1借款金额(万元)	0.486	…	0.304	0.177	…	0.517
2		X_2借款利率	0.626	…	0.411	0.201	…	0.380
3	C_2借款 者个人特征	X_5借款者年龄	1.000	…	0.600	0.850	…	0.900
4		X_7借款者学历	1.000		0.500	1.000	…	0.500
5		X_8借款者职业	1.000		0.5000	0.700	…	0.700
6		X_{10}借款者婚姻状况	1.000	…	1.000	0.500	…	1.000
7	C_3借款者财务特征	X_{11}年收入(万元)	0.700	…	0.639	0.826	…	0.639
8		X_{12}认证年收入(万元)	0.572	…	0.438	0.724	…	0.478
9		X_{14}债务收入比	0.317	…	0.217	0.142	…	0.581
10		X_{16}是否有房贷	0.500	…	0.500	1.000	…	0.500
11		X_{18}工作年限(年)	0.700		1.000	0.500	…	1.000
12		X_{19}是否拥有自有住房	1.000	…	0.500	0.500	…	1.000
13	C_4借款者信用特征	X_{21}信用评级	0.900	…	0.100	0.600	…	1.000
14		X_{22}两年内违约次数	1.000	…	1.000	0.500	…	1.000
15		X_{25}信用总额度	0.650	…	0.787	0.026	…	0.529
16		X_{26}信用使用率	0.743	…	0.031	0.865	…	0.775
17	C_5宏观经济特征	X_{28}股市收益率	0.575	…	0.648	0.615	…	0.262
18		X_{29}居民消费价格指数	1.000	…	0.651	1.000	…	0.931
19		X_{32}失业率	0.633	…	0.716	0.559	…	0.455

表8　经距离相关分析删除的评价指标

序号	距离相关系数大于0.8的一对指标				距离相关系数r_{ik}	第2次筛选删除指标
	相关指标i	D_i	相关指标k	D_i		
1	X_7借款者学历	1.597	X_8借款者职业	2.032	0.917	X_7借款者学历
2	X_{11}年收入(万元)	2.311	X_{12}认证年收入(万元)	1.842	0.841	X_{12}认证年收入(万元)
3	X_{14}债务收入比	2.017	X_{16}是否有房贷	1.610	0.932	X_{16}是否有房贷
4	X_{18}工作年限(年)	1.884	X_{19}是否拥自住房	1.716	0.924	X_{19}是否拥有自有住房
5	X_{21}信用评级	2.301	X_{25}信用总额度	1.690	0.906	X_{25}信用总额度
6	X_{22}两年内违约次数	2.174	X_{26}信用使用率	1.879	0.826	X_{26}信用使用率
7	X_{29}居民消费价格指数	1.780	X_{32}失业率	2.333	0.873	X_{29}居民消费价格指数违约次数

表9　网络借贷信用评价指标体系

序号	准则层	指标名称	指标类型
1	C_1借款标的特征	X_1借款金额(万元)	成本型
2		X_2借款利率(%)	成本型
3	C_2借款者个人特征	X_5借款者年龄	区间型
4		X_8借款者职业	定性
5		X_{10}借款者婚姻状况	定性
6	C_3借款者财务特征	X_{11}年收入(万元)	效益型
7		X_{14}债务收入比	成本型
8		X_{18}工作年限(年)	定性
9	C_4借款者信用特征	X_{21}信用评级	定性
10		X_{22}两年内违约次数	定性
11	C_5宏观经济特征	X_{28}股市收益率	效益型
12		X_{32}失业率	成本型

（三）信用评价指标体系合理性检验

根据表9的网络借贷信用评价指标体系及Logistic模型计算得到实证31000个借款者的违约概率PD_i（i=1，2，…，31000）。当取违约临界值为0.5时，即当违约概率PD低于0.5，判定贷款者为违约；不低于0.5判定贷款者为非违约。

在31000个借款者中，有4000个违约借款者与27000个非违约借款者。将4000个违约借款者的违约概率PD_i逐个与临界值0.5比较，可得实际违约借款人被判定为违约的个数DD、实际违约借款人被判定为非违约的个数DN。同理可得到实际非违约借款人被判定为违约个数ND、实际非违约借款人被判定为非违约个数NN。计算结果列入表10的数字矩阵。

表10　实际违约状态与模型判别结果

实际违约状态	模型判别结果		
	违约（1）	非违约（0）	合计数
违约	DD=3589	DN=411	$DD+DN$=4000
非违约	ND=2653	NN=24347	$ND+NN$=27000
合计	$DD+ND$=6242	$DN+NN$=24758	$TOTEL$=31000

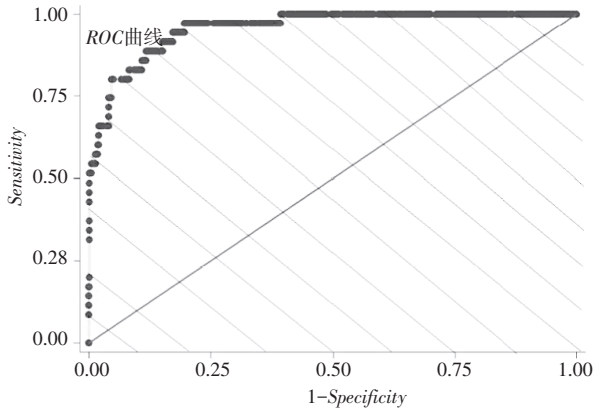

图2　检验指标体系的ROC曲线

表10的数据代入式（20），得到灵敏度为0.897；代入式（21），得到特异度为0.902，也就得到横坐标1−特异度为0.098，这样就可以确定ROC曲线上的一个点（0.098，0.897）。ROC曲线中，每取一个临界值，就得到一组灵敏度和特异度，每个（1−特异度，灵敏度）可确定一个点坐标，取不同临界值50、60…，会得到多个点，可画出ROC曲线，如图2所示。

经计算，本文构建的网络借贷信用评价指标体系对违约与非违约借款者违约状态判别精度的AUC=0.913。一般认为违约判别精度AUC超过0.8[19]时，评价指标体系的违约判别能力就较强。因此，由于本文构建的网络借贷信用评价指标体系的AUC=0.913>0.8，故认为构建的评价指标体系具有较强违约鉴别力，则评价指标体系构建合理。

五、结论与建议

（一）主要工作

本文根据K-S检验与距离相关分析相结合，筛选对借款客户违约状态甄别能力强的指标，建立了网络借贷信用评价指标体系，并通过全球最大的P2P网络借贷平台Lending Club的实际交易数据进行实证研究。结果表明：本研究评价指标体系中的借款金额、借款者职业、失业率等12个指标均对区分违约状态有显著影响。

（二）特色与创新

（1）按照K-S检验统计值愈大、其对应违约样本分布函数与非违约样本分布函数的偏离愈大，表明评价指标甄别借款客户违约状态的能力愈强，遴选能显著区分违约状态与否的评价指标，弥补现有研究不以能否区分违约状态为标准遴选评价指标的不足。

（2）通过距离相关系数反映同一准则层下两个指标间的线性与非线性关联程度，在关联程度强的一对指标中，剔除K-S检验较小、对违约状态影响较小的指标，删除了反映信息冗余指标，克服现有相关分析、因子分析等指标筛选方法仅揭示了指标间的线性关联程度，无法反映指标间非线性关联程度的弊端，拓展信用评价指标筛选方法适用范围。

（三）启示及政策建议

本文采用K-S检验与距离相关分析构建的网络借贷信用评价指标体系，实证表明：借款金额、借款者职业、失业率等12个指标均对区分违约状态有显著影响。上述研究有助于理解网络借贷违约行为及其变化规律，预测借贷违约的发生，进而在发放贷款时制定或调整相应的借贷标准，控制借贷违约的发生。

P2P网络借贷作为新型经济业态，已成为拉动国民经济的重要增长点，但由于网络借贷的违约风险较难甄别，导致当前P2P平台的相关监管存在一定程度的缺失或滞后。基于上述研究结果，本文提出如下政策建议：

（1）建议监管部门构建网络借贷违约风险评估模型，对P2P平台进行风险监测。相关监管机构可以借鉴本文评价指标筛选方法，充分利用大数据技术，识别网络借贷的违约风险，从而对P2P平台潜在的风险进行及时预判与控制。

（2）建立金融机构与P2P平台的信息共享机制，融合多源数据。由于商业银行对借款者历史违约信息了解更多，监管部门应倡导银行等机构与P2P平台加强信息共享，融合多维度信息，为准确地评估借款人违约风险提供充分的数据资源。

（四）不足之处及展望

本文主要属于应用性研究，在指标筛选方法的改进上创新有限，这也是本文的不足所在。在之后研究中，拟在信用评价指标赋权方面，按照对借款者违约状态区分能力愈强、指标赋权越大的思路，测算评价指标的权重。由于该项研究与本文的科学问题聚焦不同，因而本文未做进一步拓展，后续我们将另文专述。

参考文献

[1] 赵礼强，刘霜，易平涛. 我国P2P网络借贷平台的信用评级研究[J]. 金融理论与实践，2018(8)：58-63.

[2] 冯博，叶绮文，陈冬宇. P2P网络借贷研究进展及中国问题研究展望[J]. 管理科学学报，2017，20(4)：113-126.

[3] FRANCESCO D，FRANCESCO G，FABRIZIO C. Determinants of SME credit worthiness under basel rules：The value of credit history information[J]. PSL Quarterly Review，2014，66(264)：21-47.

[4] 谭中明，谢坤，黄丹. P2P网络借贷风险综合评价研究[J]. 农村金融研究，2017(8)：53-56.

[5] 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，2017，25(11)：12-21.

[6] 张成虎，武博华. 中国P2P网络借贷信用风险的测量[J]. 统计与信息论坛，2017，32(5)：110-115.

[7] 刘传哲，马达亮，夏雨霏. 动态异质集成信用评分模型在P2P网络借贷中的应用[J]. 金融发展研究，2018(9)：24-31.

[8] PURO L，TEICH J E，WALLENIUS H. Borrower decision aid for people-to-people lending[J]. Decision Support Systems，2010，49(1)：52-60.

[9] 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.

[10] 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[J]. 经济研究，2014(7)：125-137.

[11] EMEKTER R，TU Y，JIRASAKULDECH B，et al. Evaluating credit risk and loan performance in online peer-to-peer(P2P) lending[J]. Applied Economics，2015，47(1)：54-70.

[12] MALEKIPIRBAZARI M，AKSAKALLI V. Risk assessment in social lending *via* random forests[J]. Expert Systems with Applications，2015，42(10)：4621-4631.

[13] LIN M F，VISWANATHAN S. Home bias in online investments：An empirical study of an online crowdfunding market[J]. Management Science，2016，62(5)：1393-1414.

[14] 何光辉，杨咸月，蒲嘉杰. 中国P2P网络借贷平台风险及其决定因素研究[J]. 数量经济技术经济研究，2017(11)：44-62.

[15] 李杰，刘露. P2P网络借贷借款者违约风险影响因素研究[J]. 商业研究，2018(9)：45-54.

[16] 李延喜，孙大同，赛骞. 借款者特征对网络借贷风险的影响研究[J]. 大连理工大学学报(社会科学版)，2019，40(1)：15-22.

[17] BONFIM D. Credit risk drivers：Evaluating the contribution of firm level information and of macroeconomic dynamics[J]. Journal of Banking & Finance，2009，33(2)：281-299.

[18] LOUZIS D P，VOULDIS A T，METAXAS V L. Macroeconomic and bank-specific determinants of non-performing loans in Greece：Acomparative study of mortgage，business and consumer loan portfolios[J]. Journal of Banking & Finance，2010，36(4)：1012-1027.

[19] 王熙逸. 多元统计分析方法[M]. 上海：格致出版社，2009：178-216.

[20] SZÉKELY G J，RIZZO M L，BAKIOROV N K. Measuring and testing dependence by correlation of distance[J]. Annals of Statistic，2007，35(6)：2769-2794.

[21] 王黎明，吴香华，赵天良，等. 基于距离相关系数和支持向量机回归的PM2.5浓度滚动统计预报方案[J]. 环境科学学报，2017，37(4)：1268-1275.

[22] 迟国泰，张亚京，石宝峰. 基于Probit回归的小企业债信评级模型及实证[J]. 管理科学学报，2016，19(3)：136-156.

[23] Lendingclub. Lending Club statistics[EB/OL]. [2019-12-01]. https://www.lendingclub.com/info/downloaddata.action.

（下转第59页）

tion of social exchange and psychological contract processes in organizations[J]. Academy of Management Journal, 2008, 51(6): 1079-1098.
[59] HOFMANN D A, GAVIN M B. Centering decisions in hierarchical linear models: Implications for research in organizations [J]. Journal of Management, 1998, 24(5): 623-641.

The Construction and Verification of the Scale of Internal Innovational Entrepreneurial Environment in Chinese Enterprises

Zhang Haitao[1], Xiao Lan[2], Li Peng[3], Chen Guopeng[1]
(1. School of Management, Hubei University of Education, Wuhan 430205, China;
2. School of foreign studies, Jiangxi Agricultural University, Nanchang 330045, China;
3. Business School, Wuchang Polytechnic College, Wuhan 430202, China)

Abstract: The paper explores the test on the internal innovational entrepreneurial environment of the company based on entrepreneurial environment, and filter the items after pre-survey according to the survey of five dimensions about innovational entrepreneurial environment, then collate and analyze the original data, and carry out the exploratory factors analysis and confirmatory factors analysis of the scale, and verify the testing on innovational entrepreneurial environment of the company from the perspective of reliability and validity. Finally, the paper studies the effectiveness of the internal innovational entrepreneurship environment, As a result, the data indicate the scale of innovational entrepreneurial environment of the company including management support, resource support, compensation support, organizational culture and organizational structure has good reliability and validity. The perception of internal innovational entrepreneurship environment in enterprises has a significant positive impact on the internal innovational entrepreneurship behavior of employees. Psychological empowerment completely mediates the influence of various dimensions of the perception of internal innovational entrepreneurship environment in enterprises on the internal innovational entrepreneurship behaviors of employees.
Keywords: entrepreneurial environment; innovational entrepreneurial environment; five factors; psychological empowerment; the internal innovational entrepreneurship behavior

(上接第47页)

Construction of Evaluation Index System of Network Credit Based on Significant Differentiation of Default

Duan Chong
(School of Economics, Inner Mongolia University of Finance and Economics, Hohhot 010070, China)

Abstract: As a new model of Internet finance, online lending improves the efficiency of financial resources utilization and alleviates the difficulty of small enterprises in financing. It is of great significance for the healthy and sustainable development of Internet finance to construct a reasonable credit evaluation index system of online lending, so as to screen and prevent the potential risks of online lending in time. Based on the combination of *K-S* test and distance correlation analysis, this paper selects the indicators with strong ability to screen the default status of borrowing customers, and establishes the credit evaluation index system of online lending The empirical research on lending Club transaction data shows that not only the loan amount, loan interest rate and other characteristics of the loan object have significant correlation with the borrower's default, but also the personal characteristics such as the borrower's age, the financial characteristics such as the borrower's annual income, and the number of borrowers' default have a significant impact on the borrower's default risk. When lending funds, investors tend to favor the borrowers who are married, of moderate age, have a certain work experience, and have less historical defaults. Therefore, the risk supervision department should build a default risk assessment model of online lending, monitor the risk of P2P platform, establish a key information sharing mechanism, integrate multi-source data, and clarify the scope of review, so as to realize the healthy and orderly development of P2P network lending industry.
Keywords: network lending; credit evaluation; significant distinction of default; *K-S* test; rank correlation coefficient