基于最小二乘法的网络借贷模型

吴小英,鞠 颖*
(厦门大学信息科学与技术学院,福建 厦门 361005)

摘要:网络借贷是近年来出现的新兴事物,研究网络借贷中借款用途对借贷成功率的影响.通过使用美国最大的 P2P 网络借贷 Prosper 网站 2007 年以来的数据,以单一变量原则建立数学模型,并使用最小二乘法进行参数估计,回归实证研究表明:同等条件下,学生借款比其他种借款成功率低 3.4%,分析其原因发现学生还款率并不比其他的低,因此对学生存在直觉歧视.通过类似的方法,还发现用于汽车或者其他方面的债务更容易借到钱.同时,实证研究还得出了其他若干因素,如借款金额、利率等对借款成功率的影响.

关键词:P2P 借贷;借款用途;借款成功率
中图分类号:TP 305⁺.1 **文献标志码**:A **文章编号**:0438-0479(2012)06-0980-05

随着互联网技术的快速发展和普及,促使 P2P 网络借贷的诞生,P2P 借贷是指不需要以银行等传统金融机构为中介,借贷双方直接通过网络平台交易的无担保借贷[1-5].借款人可以低于银行贷款利息快速借到钱而放款人可获得高于银行存款利息.P2P 借贷模式在网络上发展很快.目前,美国最大的网络借贷平台是 Prosper[6],拥有超过 120 万的成员及将近 30 亿美元交易量,欧洲最大的网络借贷平台是 Zopa.国内正处于初步发展阶段,出现了宜信、拍拍贷、人人贷等网络借贷平台.

作为新的金融模式,对网络借贷已有的研究主要集中于社交关系,借款人的信用、性别、种族,以及借款的金额、利率等信息[7-13].然而,迄今为止,在国内外尚未见到对借款用途与借款成功率关系的研究报道.从直觉上看,借款用途说明了借款人将如何使用这一笔钱,其对能否借款成功应该有一定的影响.本文试图研究网络借贷平台上借款用途对借款成功率的影响,通过分析美国 Prosper 平台 2007 年以来的数据,得出不同借款用途对借款成功率的影响.

1 网络借贷数据介绍

在 Prosper 借贷交易平台,借款人发帖提交借款信息,放款人对其进行投标,如果在期限内投标金额达到所需金额则生成借贷关系.

Prosper 平台提供了借款金额、利息、借款用途,以及借贷者的属性如信用等级、住房情况、城市住址等信息.Prosper 根据借贷者的信用得分将信用等级分为 7 个等级,分别为 AA、A、B、C、D、E 和 HR.Garman 等[12]研究了不同信用等级的借贷者的借款金额、投标个数、利率等,发现信用等级大于等于 E 且利息高于 0.1 的借款,其借款成功概率较大,而等级小于 E,即使其利息高达 0.28 仍然有一大部分不能成功借款.本文选取的数据39 371个成功借贷中有 8 802 个到期未还款,17 612个已还款,12 957个正在还款中,还款率为 66.68%.

图 1 是 2007 年以来各种用途借款及还款情况,图 1(a)中可以看出学生债务的借款成功率最低,而汽车债务及其他债务的借款成功率较高.从图 1(b)中可以看出正在还款中基本没有不明确用途的借款,说明近期放款人都不愿意选择投资不明确用途的借款.从直观上可以得出投资人在选择借款时会考虑到借款用途.

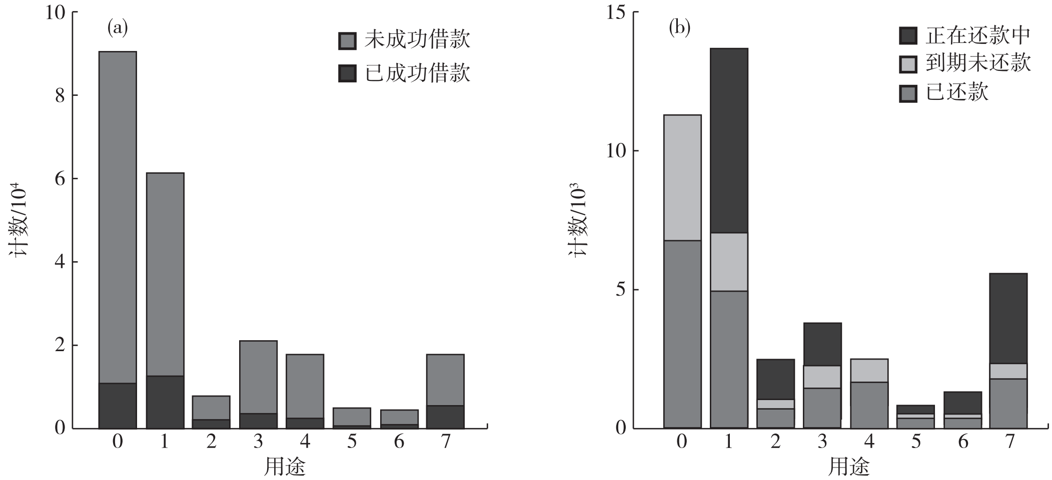
2 数学模型建立

在本文的研究中,因变量借款成功率不止受借款用途的影响,还受其他几个重要因素的影响.此时,就需要用多个影响因素作为自变量来解释因变量的变化,通过回归系数得到其对因变量的影响.

2.1 多元线性回归模型及回归系数的最小二乘估计

当随机变量 Y 的自变量 X 有 m 个,其线性回归模

收稿日期:2012-03-06
* **通信作者**:yju@xmu.edu.cn



0. 不明确用途；1. 债务巩固；2 表示提高家庭生活；3. 商业债务；4. 个人债务；
5. 学生债务；6. 汽车债务；7. 其他债务.

图 1　各种用途的借款(a)还款(b)情况图

Fig. 1　The borrow (a) and repayment (b) rate of different category

型[14]为

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m + \varepsilon, \tag{1}$$

式中 α、β_1、β_2、…、β_m 为回归系数，ε 为随机误差，通常假设 ε 遵从正态分布，$\varepsilon \sim N(0,\sigma^2)$. 回归分析的主要目标是利用观测值来估计参数，下面介绍最小二乘法进行参数估计. 令 $\hat{\alpha}$、$\hat{\beta}_1$、$\hat{\beta}_2$、…、$\hat{\beta}_m$、$\hat{\varepsilon}$、$\hat{y}$ 为参数 α、β_1、β_2 … β_m、ε 及观测值 y 的最小二乘估计，则有

$$y_k = \hat{\alpha} + \hat{\beta}_1 x_{1k} + \hat{\beta}_2 x_{2k} + \cdots + \hat{\beta}_m x_{mk} + \hat{\varepsilon_k}, \tag{2}$$

$$\hat{y}_k = \hat{\alpha} + \hat{\beta}_1 x_{1k} + \hat{\beta}_2 x_{2k} + \cdots + \hat{\beta}_m x_{mk}, \tag{3}$$

$$\hat{\varepsilon}_k = \hat{y}_k - y_k. \tag{4}$$

根据最小二乘法，$\hat{\alpha}$、$\hat{\beta}_1$、$\hat{\beta}_2$、…、$\hat{\beta}_m$ 应使得全部的观测值 y_k 与回归值 $\hat{y}_k$ 的偏差平方和 Q 最小，即求

$$Q = \sum \hat{\varepsilon}_k^2 = \sum [y_k - (\hat{\alpha} + \hat{\beta}_1 x_{1k} + \hat{\beta}_2 x_{2k} + \cdots + \hat{\beta}_m x_{mk})]^2, \tag{5}$$

的最小值. 因为 Q 是 $\hat{\alpha}$、$\hat{\beta}_1$、$\hat{\beta}_2$、…、$\hat{\beta}_m$ 的非负二次式，因此一定有最小值，可以通过 Q 对 $\hat{\alpha}$、$\hat{\beta}_1$、$\hat{\beta}_2$、…、$\hat{\beta}_m$ 求偏导，从而解出系数值.

2.2　借款数学模型的建立

有很多因素都会对借款成功与否有影响，为了验证借款用途对借款成功与否的关系，我们采取单一变量原则，建立了如下数学模型：

$$I = \alpha_1 + \gamma_1 X + \zeta_1 W + \beta_{1i} + \varepsilon_1, \tag{6}$$

其中，I 表示借款成功与否(=1 借款成功，=0 借款失败)，X 表示借款属性包括借款金额、利率等，W 表示借款人的属性包括信用等级、债务与收入比例、住房情况等. 借款用途有 0～7，这 8 个值并没有数量大小的关系，因此应作为哑元处理，用 β_{1i} 表示用途 i 的权值，在此可将 β_{17} 设为 0，其余参数与此为参照就可以估计出来. 该模型包含了此模型 X、W、β_{1i} 这些影响放款人投标的基本变量，ε_1 表示模型的误差. 通过此估计模型中的 β_{1i} 值可以得到借款用途 i 对借款成功与否的影响.

模型(6)的计算结果 I 即为二态变量，1 表示借款成功，0 表示借款失败. 但是，后面式子的结算结果，其值会介于 0 和 1 之间. 因此，本文将 I 修改为 P(其值在 0 到 1 之间)，其中 P 表示成功的概率(投款完成的百分比). 另外，模型(6)把借款用途作为一个哑元，有 $\beta_{10} \sim \beta_{17}$ 8 个值，在此可将其表示为 $\beta_{2i}\mathrm{Category}_i$，其中 $\mathrm{Category}_i$ 为 0/1 变量，由此模型(6)可变为：

$$P = \alpha_2 + \gamma_2 X + \zeta_2 W + \beta_{2i}\mathrm{Category}_i + \varepsilon_2, \tag{7}$$

其中，P 表示借款成功概率，X 仍表示借款属性包括借款金额、利率等，W 表示借贷者的属性包括借款信用等级、债务与收入比例、是否有住房等，$\mathrm{Category}_i$ 是一个哑元，$\mathrm{Category}_i$ 表示用途 i(=1 表示有用途，=0 表示没有用途)，ε_2 表示模型的误差. 由这个模型，可以得出每个用途的系数 β_{2i}，即该用途对借款成功率的影响(利/弊).

3　实验结果及讨论

本文在 Prosper 平台的数据集上使用模型(6)建模，选用借款金额、投标个数、借款利率表示模型中的 X，选用信用评分、债务与收入比、住房情况、每月还款

额表示模型中的 W,分别以借款是否成功(=1 借款成功,=0 借款失败)和投款完成百分比作为因变量 I,将借款类型 7(Category=7)的系数 β_{17} 为 0,$\beta_{10}\sim\beta_{16}$ 以此为参照,通过最小二乘估计各参数值,结果如表 1 所示. 表 1 中系数栏的估计值,如果大于 0,说明此项对借款成功率的作用为正,且其值越大说明越容易借款;若该值等于 0 说明此项对借款成功与否无影响;如果该值小于 0,说明此项对借款成功率的作用为负,且其值越小越不容易借款. 表中参数估计的标准差,其值越小说明结果越好.

表 1 模型(1)最小二乘法回归系数估计值

Tab. 1 The estimates of OLS regression coefficients by model(1)

参数	借款是否成功		投款完成百分比	
	系数	标准差	系数	标准差
截距	−0.472	0.012	−0.825	0.01
借款金额	−0.2534	0.0098	−0.2682	0.0085
投标个数	0.002	0.00001	0.002	0.000001
借款利率	0.106	0.015	0.399	0.013
信用评分	0.001	0.00002	0.002	0.00001
债务与收入比	−0.007	0.001	−0.011	0.001
住房情况	−0.004	0.002	−0.023	0.002
每月还款额	0	0.00003	0	0.00002
[Category=0]	0.007	0.004	0.01	0.003
[Category=1]	−0.006	0.004	−0.038	0.003
[Category=2]	−0.024	0.006	−0.007	0.005
[Category=3]	−0.065	0.005	−0.061	0.004
[Category=4]	−0.018	0.005	−0.042	0.004
[Category=5]	−0.051	0.008	−0.067	0.007
[Category=6]	0.003	0.007	0.012	0.006
[Category=7]	0^{a}		0^{a}	

注:a. 此参数为冗余参数,将被设为 0.

从每种借款用途的系数,可以看出该借款用途对借款成功率、投款完成百分比的影响. 观察可见,商业债务(Category=3)和学生债务(Category=5)的系数值最小,说明放款人最不愿把钱投给这两种用途的借款人. 表 2 统计了各个借款用途对应的各个属性的所有贷款、成功贷款、未成功贷款、已还款、已违约的平均值和还款率. 从表 2,可以看出用途 3 还款率相对较低,这是一种统计歧视;而用途 5 的还款率处于中上水平,这属于直觉歧视[15]. 那么,究竟是否存在歧视呢?我们通过模型(7)进一步验证.

由于借款用途 0 的个数比较少,我们只需按模型(7)设置 Category_1 到 Category_7 这 7 个 0/1 变量,所选参数 X 、W 与模型(6)一样,因变量 P 为投款完成百分比. 通过最小二乘法估计,得出各借款用途系数 β_{2i},如表 3 所示. 观察可见,学生贷款(Category=5)的系数最小且为负,该结果说明与其他借款类型相比,学生贷款在其他条件一样(X 、W 取相同值)的情况下,借款成功率比其他要低 3.4%,存在对学生贷款的歧视,而统计结果也表明借款成功率仅有 15.79%,低于平均水平. 从表 2 还可看出,学生贷款的贷款利率为 24.5%,学生贷款的还款率为 73.8%,均高于平均水平. 分析其原因可能是放款人认为学生的信用得分较低且收入不高,所以不愿借钱给学生,但事实证明学生还款率并不低,因此对学生贷款的歧视为直觉歧视. 从表 3 还可看出,汽车债务(Category=6)和其他债务(Category=7)更容易借到款,而事实上这两类借款的利率和还款率确实较高(见表 2),放款人对这两类借款的偏好可以得到统计数据的支持.

从表 1 可见,借款金额、债务收入比的系数为负,说明这些值越大,借款成功概率越低. 这二者的原因是显而易见的:借款金额越大,越难借到钱;债务收入比越大,说明借款人负债较多,还款能力较小,放款人不愿投资这样的借款. 住房情况的系数很小,说明对借款成功概率影响并不大;每月还款额的系数为 0,说明对借款成功率没有影响. 借款利率、信用评分的系数为正,说明借款利率越高,借款成功概率越高,这与文献[15]的研究结果一致;信用得分是由官方公布的数据结合 Prosper 平台上的历史记录得出,信用得分越高说明信用越好,还款概率越大,这一结论符合 Katherine 等得出放款人愿意选择投钱给信用等级 C 及以上的研究结果[9].

4 结 论

本文通过数学模型对网络借贷平台 Prosper 上的借款用途进行分析. 模型以单一变量原则,通过最小二乘法估计系数,从得到的系数估计值可以看出借款用途对借款成功率的影响:学生债务较难借到钱而汽车债务相对容易借到钱,而对学生债务的统计表明学生债务的借款利率和还款率并不低(高于平均水平),由此推断放款人对学生债务存在直觉歧视. 模型还得到借款金额、借款利率、信用评分等对借款成功率的影响,结果与已有文献的结论一致.

表 2 不同用途各个参数的平均值

Tab. 2 Different parameters′ mean value of different categories

参数	0	1	2	3	4	5	6	7
还款率	0.604	0.704	0.658	0.658	0.676	0.738	0.798	0.777
信用评分								
所有借款	581.3	642.6	686.1	662.8	594.9	620.9	655.2	659.0
成功借款	646.7	681.6	707.4	705.1	656.3	663.3	687.0	690.3
未成功借款	568.8	626.7	674.0	650.5	577.3	609.9	639.1	642.0
已还款	670.0	674.5	709.2	708.3	670.4	664.8	687.5	689.1
违约	615.8	641.7	683.5	683.4	629.6	633.9	650.7	646.2
借款利率								
所有借款	0.175	0.230	0.227	0.233	0.219	0.245	0.231	0.249
成功借款	0.176	0.208	0.203	0.196	0.181	0.205	0.208	0.219
未成功借款	0.175	0.237	0.238	0.241	0.225	0.253	0.241	0.262
已还款	0.156	0.187	0.176	0.166	0.163	0.197	0.177	0.195
违约	0.207	0.227	0.209	0.207	0.217	0.234	0.205	0.250
借款金额								
所有借款	8226	7213	7500	10080	5169	5008	5191	7504
成功借款	70887	63419	64536	7888	4556	3514	4282	4591
未成功借款	8379	7446	7956	10532	5263	5289	5550	5734
已还款	6633	6055	6096	7751	4528	3595	4269	4036
违约	7756	5849	7681	9778	4616	3664	4152	4204
债务与收入比								
所有借款	0.665	0.369	0.272	0.37	0.398	0.415	0.268	0.322
成功借款	0.440	0.264	0.217	0.256	0.235	0.268	0.204	0.225
未成功借款	0.696	0.399	0.300	0.399	0.425	0.447	0.296	0.368
已还款	0.381	0.254	0.199	0.255	0.223	0.277	0.183	0.211
违约	0.531	0.300	0.235	0.307	0.258	0.270	0.196	0.255
住房情况								
所有借款	0.57	0.46	0.69	0.51	0.34	0.26	0.34	0.41
成功借款	0.45	0.51	0.74	0.56	0.42	0.26	0.39	0.44
未成功借款	0.32	0.42	0.65	0.48	0.33	0.22	0.29	0.37
已还款	0.44	0.47	0.73	0.56	0.41	0.25	0.36	0.45
违约	0.45	0.50	0.73	0.59	0.42	0.27	0.31	0.41
投标个数								
所有借款	0.01	21.83	4.43	8.08	3.26	0.91	1.56	6.94
成功借款	0.01	65.82	12.03	19.82	9.75	2.6	4.41	19.18
未成功借款	0	7.28	1.92	4.20	1.12	0.35	0.62	2.91
已还款	0	71.85	11.37	24.3	17.82	3.98	4.77	18.37
违约	0	62.85	11.65	33.32	18.64	3.36	2.93	11.36

注:0～7 的表示见图 1,表 3 同.

表 3 模型(7)最小二乘法回归的系数估计

Tab. 3 The estimates of OLS regression coefficients by model(7)

用途	1	2	3	4	5	6	7
系数	−0.009	0.027	−0.014	−0.026	−0.034	0.040	0.037

参考文献:

[1] 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):101-104.

[2] 王继晖,李成. 网络借贷模式下洗钱风险分析及应对[J]. 金融与经济,2011(9):9-11.

[3] 陈盛东. 依法管贷后记:借款用途莫虚设[J]. 中国农村信用合作,2001(1):32-33.

[4] 邢增艺,王艳. 网络借贷:微型金融发展新趋势[J]. 前沿,2010(23):109-111.

[5] 王艳,陈小辉,邢增艺. 网络借贷中的监管空白及完善[J]. 当代经济,2009(24):46-47.

[6] Prosper Company. Prosper marketplace[EB/OL][2012-09-10]. http://www.prosper.com.

[7] Berger S, Gleisner F. Emergence of financial intermediaries in electronic markets: the case of online P2P lending[J]. Business Research, 2009, 2(1):39-65.

[8] Krumme K, Herrero-lopez S. Do lenders make optimal decisions in a peer-to-peer network? [C]//Web Intelligence and Intelligent Agent Technologies. Milan, Italy: IEEE, 2009:124-127.

[9] Lauri P, Teich J E, Wallenius H, et al. Borrower decision aid for people-to-people lending[J]. Decis Support Syst, 2010, 49(1):52-60.

[10] Luo C, Xiong H, Zhou W, et al. Enhancing investment decisions in P2P lending: an investor composition perspective[C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2011:292-300.

[11] Barasinska N, Schäfer D. Are women more credit-constrained than men-evidence from a rising credit market [C]//Advisory and Steering Committee Meeting. Berlin, Germany: FINESS, 2010:1-27.

[12] Hampshire S R, Krishnan R. Asearch theoretic model of person-to-person lending[EB/OL][2012-09-10]. http://heinz.cmu.edu/research/244full.pdf.

[13] 刘严. 多元线性回归的数学模型[J]. 沈阳工程学院学报:自然科学版,2005,1(增刊 1):128-129.

[14] Li Siming, Qiu Jiaxian, Lin Zhangxi, et al. Do borrowers make homogeneous decisions in online P2P lending market? an empirical study of PPDai in China[C]//2011 8th International Conference on ICSSSM. Tianjin, China: IEEE, 2011:1-6.

[15] Becker G S. The economics of discrimination[M]. Chicago: University of Chicago Press, 1971.

Online Lending Model Based on Ordinary Least Square

WU Xiao-ying, JU Ying*

(School of Information Science and Technology, Xiamen University, Xiamen 361005, China)

Abstract: P2P online lending is an emerging economic lending model. In this work we study the influence of the purpose of lending on success rate at online lending marketplace. Using the data of American largest P2P lending market Prosper since 2007, we establish mathematical model, and employ ordinary least square, univariate multiple regression. The experimental results show that, the successful rate of student loan is 3.4% lower than the others, ceteris paribus. The reason is not due to low student repayment, but because of taste-based discrimination. On the other hand, for auto loan or other loan, it is easier to get fund successfully. Meanwhile, we also investigate the influence of other factors, such as loan amount and rate.

Key words: P2P lending; lending purpose; successful fund rate