DOI: 10.19634/j.cnki.11-1403/c.2018.04.001

基于 Logistic-DEA 的互联网金融贷款产品有效客户识别

□ 刘冰清[1] 卢子芳[1] 朱卫未[1] 尹相菊[2]

(1.南京邮电大学 管理学院，江苏 南京 210000；2.华东师范大学 经济与管理学院，上海 200333)

[摘要] 收集互联网金融贷款产品用户的海量多维数据，利用数据包络分析(DEA)对数据进行一些指标预处理，在传统逻辑回归模型中增加 DEA 效率值，提高模型准确率。结果证明，互联网贷款产品的客户的短信回应率得到了显著的提高，使得此款互联网贷款产品的潜在客户得到有效识别，达到精准营销与增加收入的目的。

[关键词] 互联网贷款产品；互联网金融；精准营销；DEA

[中图分类号]F832 [文献标识码]A [文章编号]1003-1154(2018)04-0001-04

一、引 言

精准营销是近年来互联网金融产品客户有效识别的重要途径之一。精准营销指的是在存储海量客户数据的基础上，依托现代信息技术手段建立个性化的顾客沟通服务体系，实现公司可度量的低成本扩张之路。对于互联网消费金融贷款产品的有效客户识别，实际上是对网贷产品精准营销的推动，可以去除传统金融产品的大量弊端，为新兴网贷金融产品开创新纪元。

众所周知，数据库中存有非常多维度的指标，例如客户基本属性，用户金融属性，用户浏览行为轨迹等等，因此，合理分析这些数据是急需的，任何试图主观对数据进行预测的人都可能导致严重的偏见，因为关于考虑哪些关键因素的选择是十分困难的。因此，本文首先对复杂的数据进行合理的归类分析，再选用数据包络分析法以及其他合理方法对数据进行预处理，排除仅可使用原始数据的情况，增添新的有效数据值，最后用逻辑回归模型对数据集进行精准预测，达到有效客户识别的目标。

谢平和邹传伟[1]提出了互联网金融模式具有一定市场竞争力，并在优化资源配置和降低借贷成本方面优势十分明显。牛新庄[2]表示互联网金融需要以客户为中心，借助各种数据分析客户行为，准确切入场景。单纯的依靠产品本身来实现差异化已经越来越困难，对于贷款产品的客户准入标准难以统一与规范，难以进行准确高效公平的潜在客户抉择。同时，对于贷款产品的额度问题，客户经理也是无法根据每个客户的资质准确衡量。因此，利用信用评分方法以及大数据挖掘对选择有效客户识别显得尤为重要。此外，只有将产品与场景以及各客户需求紧密结合，才能被特定的客户所接受。互联网营销使得传统的“一对多”粗放模式转变为“一对一”精准营销。基于客户的价值认同开展价值营销，对目标客户进行细分，同时提供特定的金融服务，实现营销的精准投放，提供有效客户的识别。同时，沈金波和吴红[3]指出商业银行利用大数据技术，通过分析客户在网络银行、手机银行和电话银行上操作所留下的信息数据，可以逐步了解到每一名客户的特征属性，进而将产品精准定向的推送给每一个客户，提高客户的满意度。随后，赵毅[4]强调为了适应互联网金融产品的发展，除了布局大数据基础设施平台，也要积极建立大数据应用思维，通过本行内海量数据的采集、加工、存储、挖掘到基于细分场景的应用，通过实施客户画像、用户行为分析、精准营销等方法，积极接纳智能金融的理念，实现互联网消费金融产品的有效客户识别。

为了实现互联网金融贷款产品的精准营销，很多国内外学者采用了不同的方法，尤其是在信用评分领域。如 Wiginton[5]第一次把逻辑回归方法应用于信用评分领域。Carrier 和 Povel[6]在 2003 年提出回归是一种用于描绘每个数据对象的真正价值，提供预测值的统计估计技术。Koh 等[7]认为逻辑回归方法可代替判别分析方法，因为逻辑回归相对于判别分析要

[基金项目] 国家自然科学基金(71771126，71301080)；2017 年江苏省高校哲学社会科学优秀创新团队培育点(2017ZSTD022)；江苏省社会科学基金(17GLB013)。

求相对放松，即逻辑回归并不要求因变量和自变量之间是线性的，且不需要一组正态分布的变量。

数据包络分析（DEA）可用于探究将管理科学、数学、数学经济学和运筹学交叉起来的全新领域，是Chames等[8]于1978年提出的。DEA是利用多个输入和输出来衡量每个决策单元之间的相对效率。所以，在互联网消费金融贷款产品有效客户识别的过程中，DEA方法可以将每个客户作为每个决策单元，获得其效率值，而不是对客户的传统分类进行研究。Eddie等[9]为了解决项目融资中所涉及到贷款申请人的信用评估以及如何识别有效用户的问题，采用了DEA替代了传统的信用评分模型，介绍了数据包络分析方法的优点自动生成相对权重，不同于传统回归模型。占治民和罗剑朝[10]选择Logitic-DEA组合，这是基于调研实情、数据实况和复合优化而成的新尝试。思路是：在逻辑回归基础上，找到影响风险控制相关评估的主要因素，再利用研究成果选择超效率DEA进行农地承包经营权抵押贷款风险控制效果评估。Misiunas等[11]提出运用DEA做数据预处理，与神经网络模型结合，提出DEANN模型，结果表明此法对器官功能状态的预测准确率有明显的提升效果。

本文创新之处，是将DEA作为对数据做预处理方法之一，与原始数据结合，带入逻辑回归模型，对互联网贷款产品的客户进行有效的识别，实现低成本高收益的精准营销，为挖掘互联网金融产品的潜在客户提供新的视角。

二、数据准备

本文使用某互联网金融科技公司真实的信贷数据为分析对象。该公司主要经营小额贷款、线上理财、线上保险等各类互联网金融产品，拥有强大的服务网络。本文要开发的有效客户识别模型就以该公司在2016年上线的一款互联网金融产品为研究对象。该款产品的功能是为客户提供小额消费贷款，整个申请过程实现了全程自动化，无人工干预，客户只需要通过相关APP提交申请信息、验证个人资料、即可在线完成授信审批。在成功获取申请额度后，贷款将汇入客户指定的银行卡中，客户根据选择的还款期数如期还款即可。对于线上贷款产品，最大的风险就是客户违约，虽然有个人征信报告不良记录，但是对于公司来说，是一笔损失。所以，对于线上贷款产品而言，准确的识别有效客户，不仅能降低不良还款率，还能减少运营成本。

（一）数据搜集

本文研究的真实数据是由一家知名金融科技公司所提供，此公司是某集团下一个子公司，所以除了子公司数据，还可以获取到集团内部甚至外部渠道的多维度数据信息。数据中主要包括客户的基本属性、浏览属性、交易属性、问卷评估信息、注册登录信息和加挂属性。本文的数据集由205 995位客户的144项原始指标组成。根据前期做的实际营销结果，选择5 999位回复YES客户，199 996位未回复客户为本次研究对象。

（二）数据清洗

观察本数据集发现，一部分指标空值较多，还有一部分指标无太大意义，相似程度较大。首先，观察每个指标的数据饱和度，剔除空值小于20%的指标。其次，根据指标属性判别，选用众数或者中位数填补缺失值。例如，信用评价分数，贷款评价分数，基金评价分数等指标采用众数的方法。除此之外，对于一个月内最频繁登陆渠道则采用中位数的方法填充。随后，转变数据类型。例如，注册渠道、性别、居住城市等字符型数据采用woe转为数值型，注册日期等日期时间型指标用常用方法转为数值型。最后，剔除一些相似的、无区分度、无差异化的指标。

（三）数据预处理与选取

先运用WOE，将众多字符型指标变为可计量的数值型。WOE越大，则表示数值差异越大，区分度较好。对于数据选取，为了保证回YES客户的样本比例，5 999回YES用户全部保留，在所有未回YES的客户中，随机选择77%的用户及其指标数据作为训练集，剩余23%的用户及其指标数据作为测试集，即5 999个回YES的客户以及45 990个未回YES的测试集总量。在模型中中设置参数test_size=0.2，即代表随机80%的用户为训练数据，剩余20%的用户为测试数据。表1则为模型后检测出的训练集和测试集的人数分布情况。

表1 训练集和测试集的人数分布情况

	总量		训练集		测试集	
	数量	占比	数量	占比	数量	占比
0	45 990	88.46%	36 790	88.49%	9 200	88.34%
1	5 999	11.54%	4 785	11.51%	1 214	11.66%
Total	51 989	100%	41 575	79.97%	10 414	20.03%

三、Logistic-DEA模型构建

（一）逻辑回归模型构建

设因变量Y是一个二元分类变量，其取值为$Y=1$（“回YES”）和$Y=0$（“未回YES”），影响Y取值的i个自变量分别为x_1，x_2，…，x_i，设$P(Y=1)=p$，$P(Y=0)=1-p$，即p为客户回YES的概率，$1-p$为客户不回YES的概率，则逻辑回归模型为：

$$f(x)=\frac{\exp(\beta_0+\sum\beta_iX_i)}{1+\exp(\beta_0+\sum\beta_iX_i)} \quad (1)$$

构建逻辑回归模型后，用 Log 似然函数(2)求解，可得(3)

$$l(p) = \ln(L(p)) = \ln\{\prod_{i=1}^{n} P(Y = y_i)\} = \sum_{i=1}^{n} y_i \ln(p) + (1 - y_i)\ln(1 - p) \quad (2)$$

$$l(\beta) = \sum_{i=1}^{n} (y_i(\beta_0 + \sum \beta_i x_i) - \ln(1 + e^{\beta_0 + \sum \beta_i x_i})) \quad (3)$$

为了能自动剔除不重要的变量，防止过拟合，把数据清洗后的所有指标变量放入逻辑回归模型训练并加入 L2 惩罚性(4)，使其对数据维度和模型效果进行平衡，起到泛化效果，最终获取 17 个指标变量的相关系数和模型预测准确率。

$$l(\beta) = \sum_{i=1}^{n} (y_i(\beta_0 + \sum \beta_i x_i) - \ln(1 + e^{\beta_0 + \sum \beta_i x_i})) + \lambda \| \beta \|_2 \quad (4)$$

(二)DEA 模型构建与结果

标准 CCR 模型是第一个也是应用最为广泛的 DEA 模型，本文选择标准 CCR 模型，有 i 个输入指标，r 个输出指标的 n 个相互独立的决策单元，将每个客户作为每一个决策单元，分别计算其效率值，表达式如下：

$$\min \theta$$

$$\text{s.t.} \begin{cases} \sum_{j=1}^{n} x_{ij}\lambda_j \leqslant \theta x_{i0}, & (i = 1,2,\cdots,m) \\ \sum_{j=1}^{n} y_{rj}\lambda_j \geqslant y_{r0}, & (r = 1,2,\cdots,s) \\ \lambda_j \geqslant 0, & (j = 1,2,\cdots,n) \end{cases} \quad (5)$$

式中 θ 记为效率指数是决策变量，λ_j 为输入指标和输出指标系数。x_{ij} 记为第 j 个评价对象的第 i 个输入指标；y_{rj} 记为第 j 个评价对象的第 r 个输出指标。Eddie[9]指出数值越大越好的放入产出指标，数值越小越好的放入投入指标。根据逻辑回归得出的指标相关性，再根据日常业务中的经验总结，最终选择 3 个输入指标，6 个输出指标，见表 2。

表 2 DEA 模型输入和输出指标

指标	描述
输入指标	
bind_saving_cnt	存款加挂账户数
bind_ins3_cnt	产险加挂账户数
propensity_fund_decile	基金倾向分位区间
输出指标	
bind_car_cnt	车加挂账户数
propensity_smallpur_decile	小消倾向分为区间
bind_consume_cnt	积分账户加挂数
propensity_credit_decile	信用卡倾向分位区间
bind_fin_cnt	金融加挂账户数
day_30_cnt	过去 30 天登陆次数

表 3 为原先逻辑回归结果与加入 DEA 效率值的逻辑回归结果。DEA 效率值的相关性为负相关，意味着，效率值越小的用户越可能申请此贷款产品，与实际情况相符。

表 3 是否加入 DEA 效率值的逻辑回归相关性前后对比

指标	原相关性	新相关性
sclsmallpur_score	5.939 037 08	5.953 849
woe_bind_car	1.179 900 94	1.181 189
woe_gender	1.398 018 58	1.406 521
woe_bind_consume1	0.726 405 68	0.727 095
woe_bind_consume3	1.909 773 55	1.911 036
woe_logon_1month_mostfreq_source	0.592 479 03	0.592 292
woe_credit_decile	0.379 945 04	0.378 193
woe_bind_fin	0.190 725 09	0.194 256
woe_reg_source	0.264 226 89	0.263 46
woe_bind_ins1	0.462 673 98	0.465 132
woe_loan_decile	0.213 685 62	0.216 983
woe_bind_loan	0.344 057 97	0.344 892
woe_fund_decile	−0.076 285 26	−0.083 87
woe_reg_web_app_ind	0.212 445 43	0.213 661
woe_bind_ins3	−1.251 496 6	−1.262 16
woe_bind_saving1	−0.040 552 51	−0.042 43
woe_bind_other	0.301 709 14	0.301 294
DEA_socre		−0.039 06

四、验证与结果分析

文中所有模型均在 Python 中进行。结果得出，未加入 DEA 效率值时，模型的预测准确率输出为 78.260 2%，加入 DEA 效率值后的模型预测准确率，输出为79.591 0%，由此可见，DEA 效率值能够显著性提高模型的准确率，比加入其余原始指标更有效。

为了能够直观的看出模型效果，将最终所有客户的分数转化为分位区间内并进行 Kolmogorov-Smirnov 测试，KS 指标是衡量回应客户和未回应客户的累计分布比例之间距离最大的差距。首先按照样本的评分从大到小进行排序，然后计算每一个分位点段下好坏样本的累计占比。二者之间的距离越大，则 KS 指标值越高，说明该模型区分好坏客户的能力强。在实际业务中，KS 小于 20%则代表模型的准确性较差；KS 介于 20%到 30%，则模型区分效果一般；KS 介于 30%到 45%之间说明模型效果很强。此模型做出的 KS 值表示，模型预测效果很强(图 1)。

见下表 4、5 可得，在训练集中，在前三个分位区间，可以命中 78.7%有效潜在客户，即可能回 YES 的。在测试集中，在前三个分位区间，可以命中 77.1%有效潜在客户。换句话说，利用此模型，可以使用很少的短信营销成本获取 75%以上的有效潜在用户，达到有效客户识别的目的，减少公司运营成本，

取得较高收益。

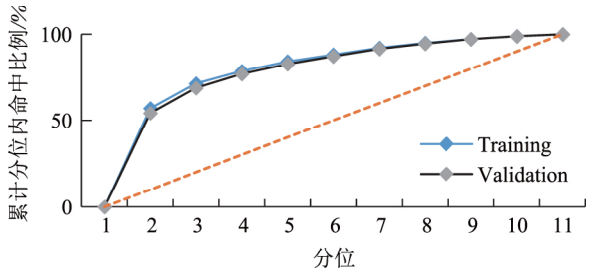

图 1　Logistic-DEA 模型中的 KS 值

表 4　训练集的预测结果及验证

十分位	分位人数	命中人数	占比	命中率	分位内命中比例	累计分位内命中比例	K. S. value
1	3 679	2 726	0. 1	74. 10%	56. 97%	56. 97%	53. 99%
2	3 679	695	0. 1	18. 89%	14. 52%	71. 49%	59. 19%
3	3 679	345	0. 1	9. 38%	7. 21%	78. 70%	55. 99%
4	3 679	266	0. 1	7. 23%	5. 56%	84. 26%	50. 88%
5	3 679	193	0. 1	5. 25%	4. 03%	88. 30%	44. 02%
6	3 679	178	0. 1	4. 84%	3. 72%	92. 02%	36. 80%
7	3 679	138	0. 1	3. 75%	2. 88%	94. 90%	28. 62%
8	3 679	110	0. 1	2. 99%	2. 30%	97. 20%	19. 77%
9	3 679	74	0. 1	2. 01%	1. 55%	98. 75%	10. 05%
10	3 679	60	0. 1	1. 63%	1. 25%	100. 00%	0. 00%
总计	36 790	4 785	1	13. 01%	100. 00%		

表 5　测试集的预测结果及验证

十分位	分位人数	命中人数	占比	命中率	分位内命中比例	累计分位内命中比例	K. S. value
1	920	659	0. 1	71. 63%	54. 28%	54. 28%	51. 02%
2	920	177	0. 1	19. 24%	14. 58%	68. 86%	56. 29%
3	920	100	0. 1	10. 87%	8. 24%	77. 10%	54. 26%
4	920	69	0. 1	7. 50%	5. 68%	82. 78%	49. 29%
5	920	54	0. 1	5. 87%	4. 45%	87. 23%	42. 89%
6	920	53	0. 1	5. 76%	4. 37%	91. 60%	36. 40%
7	920	36	0. 1	3. 91%	2. 97%	94. 56%	28. 30%
8	920	30	0. 1	3. 26%	2. 47%	97. 03%	19. 62%
9	920	22	0. 1	2. 39%	1. 81%	98. 85%	10. 19%
10	920	14	0. 1	1. 52%	1. 15%	100. 00%	0. 00%
总计	9 200	1 214	1	13. 20%	100. 00%		

五、结　论

根据以上研究结果，可看出由于 DEA 指标的加入，使得逻辑回归模型的准确率得到显著的提高。可见，DEA 指标考虑进互联网金融产品的有效客户识别预测模型中是有意义的。模型可以帮助公司通过过滤无效客户减少成本和挖掘潜在客户。这种精准营销的方式不仅可以用在互联网金融产品，也可以广泛用在各大行业。随着大数据的发展，数据的获取方式和维度将会越来越广泛与多样化，加入更多可靠的指标，例如客户的人脉关系信息，客户的消费能力等，从而使模型更加合理，更加全面。

除此之外，建模方法上可以改进。虽然基于逻辑回归得到的评分模型效果较好，但是在预测精度方面，人工智能方法效果更佳。因此，随着人工智能方法的完善和实践的深入，可以利用人工智能方法建立评分模型，提高模型的预测精准度。回归到数据本身，最大的挑战是保持这些商业记录随着关系的发展和扩大而不断发展。只有源数据非常好，再结合有效的数学模型，就可以利用大数据实现精准营销。□

通讯作者：朱卫未，E-mail：kirbyzhu@163. com。

［参考文献］

［1］谢平，邹传伟. 中国 P2P 借贷服务行业白皮书［M］. 北京：中国经济出版社，2014.

［2］牛新庄. 对互联网金融产品创新的思考［J］. 银行家，2015(03)

［3］沈金波，吴红. 大数据环境下的银行网络营销策略研究［J］. 电子商务，2015(12).

［4］赵毅. 借力大数据平台实现科技金融创新［J］. 金融电子化，2017 (02).

［5］Wiginton J C. A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior［J］. Journal of Financial & Quantitative Analysis，1980，15(03)：757-770.

［6］Carrier C C，Povel O. Characterising Data Mining Software［J］. Intelligent Data Analysis，2003 (07)：181192.

［7］Koh H C，Tan W C，Goh C P. Credit Scoring Using Data Mining Techniques［J］. Singapore Management Review，2004，26(02)：25-47.

［8］Chames A，Cooper W W，Phodes E. Easuring the Efficiency of DMU［J］. European Journal of Operational Research，1978，11(06)：429-444.

［9］Eddie W L. Cheng，Yat Hung Chiang，Bo Sin Tang. Alternative Approach to Credit Scoring by DEA：Evaluating Borrowers with Respect to PFI Projects［J］. Building and Environment，2007，42 (04)：1752-1760.

［10］占治民，罗剑朝. 基于 Logistic-DEA 的农村土地承包经营权抵押贷款试点风险控制效果评估［J］. 武汉大学学报(哲学社会科学版)，2016 (05).

［11］Misiunas N，Oztekin A，Chen Y，et al. DEANN：A Healthcare Analytic Methodology of Data Envelopment Analysis and Artificial Neural Networks for the Prediction of Organ Recipient Functional Status［J］. Omega，2016，58(15)：46-54.