DOI: 10.16661/j.cnki.1672-3791.2019.03.255

基于逻辑回归的商业银行客户信用评级研究①

郝婷婷 俞俊杰 陈燕
(山东省农村信用社联合社 山东济南 250014)

摘 要：客户信用评级是衡量客户偿债能力和偿债意愿的重要标准，是指导商业银行及各金融机构信用风险管理的重要手段。各商业银行建立自己内部合适有效的信用评分模型对于管控风险、降低损失尤为重要。逻辑回归分析是一种广义的线性回归分析模型，在数据挖掘、经济预测等领域有着广泛的应用。该文采用统计学建模方法，将逻辑回归模型应用于农村商业银行客户信用评级，构建了基于逻辑回归的农村商业银行客户信用评级模型，对贷款客户的信用等级进行评价。该方法能够帮助银行有效区分违约客户，比较准确地量化客户违约风险，同时为商业银行及其他金融机构开展信用风险管理提供参考。

关键词：客户信用评价 逻辑回归 违约风险
中图分类号：F830.33 **文献标识码**：A **文章编号**：1672-3791(2019)01(c)-0255-02

随着经济全球化的发展，金融市场的波动性日渐加剧，商业银行作为金融机构体系中最为核心的部分，所面临的金融风险具有越来越多的不确定性。与此同时，信贷产品不断丰富，各类信用消费纷纷涌现，商业银行的信用风险问题越来越突出，风险暴露越来越严重。然而日益激烈的市场竞争要求各商业银行及金融机构更好的管控信贷风险，同时作为经济主体，如何降低自身损失，获取最大利润，是信用决策者一直关注的问题。

信用评分模型通过挖掘银行系统现有信用样本的有效信息，对影响客户信用评级中的各种因素进行综合考量，从而对信用风险做出预测判断。

农村商业银行由于其服务对象的特殊性等一些原因，客户信用评级难以估量，用传统的风险评级方法难以达到满意的结果。因此，建立自己内部合适有效的信用评分模型，合理评价客户信用等级，预测贷款客户是优良顾客还是违约顾客，对于管控信贷风险、提高银行效益、降低银行损失显得尤为重要。但从相关研究看来，我国还缺乏一套成熟完整的适用于农村商业银行的信用评级模型，因此，文章拟通过逻辑回归方法，建立适用于农村商业银行的信用评级模型，为农村商业银行更有效地进行信用评价提供新的评级依据。

1 逻辑回归理论

逻辑回归[1, 2]是一种广义的线性回归分析模型，在数据挖掘、疾病自动诊断、经济预测等各个领域有着广泛的应用。

考虑具有P个指标变量构成的向量$x^{'}=(x_1,x_2\cdots x_p)$，假设用$Y$表示客户信用状况这一事件，则$Y$=1表示客户信用良好，$Y$=0表示信用情况恶劣。设条件概率$P(Y=1|x)=p$为根据观测量相对于某事件发生的概率，则逻辑回归模型可表示为：

$$P(Y=1|x)=\pi(x)=\frac{1}{1+e^{-g(x)}} \tag{1}$$

其中$g(x)=\beta_0+\beta_1x_1+\cdots+\beta_px_p$，$\pi(x)=\frac{1}{1+e^{-g(x)}}$是Logistic函数。

定义事件不发生的条件概率为：

$$P(Y=0|x)=1-P(Y=1|x)=\frac{1}{1+e^{g(x)}} \tag{2}$$

那么，事件发生与事件不发生的概率之比为：

$$\frac{P(Y=1|x)}{P(Y=0|x)}=\frac{p}{1-p}=e^{g(x)} \tag{3}$$

这个比值称为事件的发生比，简称为odds。因为0<P<1,故odds>0。对odds取对数，即得到线性函数，

$$\ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1x_1+\cdots+\beta_px_p \tag{4}$$

其中β_1, β_2, ..., β_p为待估参数。

利用最大似然估计来测算偏回归系数β_0, β_1, ..., β_p。其基本思想是先建立似然函数与对数似然函数，求使对数似然函数最大时的参数值，其估计值即为最大似然估计值。

设Y是0-1变量，$(x_1,x_2\cdots x_p)$是与Y相关的自变量，m组观测数据为$(x_{i1},x_{i2}\cdots x_{ip},y_i)$，$i$=1,2, ..., m, 则$(y_1,y_2,\cdots,y_m)$的似然函数为：

$$K=\prod_{i=1}^{m}p(y_i)=\prod_{i=1}^{m}p(x_i)^{y_i}[1-p(x_i)]^{1-y_i} \tag{5}$$

建立对数似然函数：

$$\ln K=\sum_{i=1}^{m}[y_i(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots\beta_px_{ip})-\ln(1+e^{\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots\beta_px_{ip}})] \tag{6}$$

上式分别对β_0, β_1, ..., β_p求偏导，求得使得对数似然函数最大的$\hat{\beta_0},\hat{\beta_1},\cdots,\hat{\beta_p}$，即为逻辑回归系数$\beta_0$, β_1, $\cdots$, β_p的估值。

2 信用评分模型构建

农村商业银行的贷款农户对象以小微型企业、个体工商户、个体农户为主。文章以个体农户为主要研究对象建立基于逻辑回归模型的客户信用评级模型。

2.1 评价指标选取

影响客户信用评级的因素有很多，调查发现，客户年龄、婚姻状况、受教育程度、收入等基本信息在客户信用评级中起着非常重要的作用[3]。选择合适的评级指标是信用评级模型准确有效的关键，同时选择评价指标体系也要遵循全面、合理、数据易得的原则。商业银行一般选取月收入作为贷款审批的重要因素，但个体农户作为特殊的客户群体，收入来源不固定，月收入不稳定，因此文章选取收入来

①作者简介：郝婷婷（1990—），女，汉族，山东济南人，硕士，主要从事核心银行系统贷款研究工作。
俞俊杰（1987—），男，汉族，山东青岛人，硕士，主要从事核心银行系统贷款研究工作。
陈燕（1989—），女，汉族，山东德州人，硕士，主要从事核心银行系统贷款研究工作。

源、年收入作为衡量客户还款能力的重要指标。综上，文章选取指标客户年龄、受教育程度、收入来源、年收入、婚姻状况、贷款金额、还款方式、担保方式、贷款年限。

2.2 指标数据量纲处理

（1）数值型指标。

数值型指标即取值用数值大小衡量的指标，例如客户年龄、年收入、贷款期限、贷款金额等指标。采用如下公式进行标准化。

均值标准差模式：新数据=（原数据-均值）/标准差

（2）字符型指标。

对于取值有多种情况的指标，如婚姻状况、受教育程度、收入来源、还款方式、担保方式等指标，引入虚拟变量，事先把这些字符型指标重新编码，生成多个二分类虚拟变量。

婚姻状况（以未婚单身为基准）：

$$\text{婚姻状况1}=\begin{cases}1 & ，已婚\\ 0 & ，其他\end{cases}\quad \text{婚姻状况2}=\begin{cases}1 & ，离异\\ 0 & ，其他\end{cases}$$

受教育程度（以初中及以下为基准）：

$$\text{受教育程度1}=\begin{cases}1 & ，高中\\ 0 & ，其他\end{cases}\quad \text{受教育程度2}=\begin{cases}1 & ，大学\\ 0 & ，其他\end{cases}$$

$$\text{受教育程度3}=\begin{cases}1 & ，研究生\\ 0 & ，其他\end{cases}$$

收入来源（以务农为基准）：

$$\text{收入来源1}=\begin{cases}1 & ，经商\\ 0 & ，其他\end{cases}\quad \text{收入来源2}=\begin{cases}1 & ，打工\\ 0 & ，其他\end{cases}$$

还款方式（以等额还款为基准）：

$$\text{还款方式1}=\begin{cases}1 & ，期末本息一次付清\\ 0 & ，其他\end{cases}\quad \text{还款方式2}=\begin{cases}1 & ，定期结息，到期还本\\ 0 & ，其他\end{cases}$$

$$\text{还款方式3}=\begin{cases}1 & ，按计划表还本付息\\ 0 & ，其他\end{cases}\quad \text{还款方式4}=\begin{cases}1 & ，期初付息\\ 0 & ，其他\end{cases}$$

担保方式（以第三方担保为基准）：

$$\text{担保方式1}=\begin{cases}1 & ，信用\\ 0 & ，其他\end{cases}\quad \text{担保方式2}=\begin{cases}1 & ，抵押\\ 0 & ，其他\end{cases}\quad \text{担保方式3}=\begin{cases}1 & ，质押\\ 0 & ，其他\end{cases}$$

2.3 样本数据预处理

在进行数据挖掘前，为提升挖掘的效率，需要进行数据的预处理，包括数据质量的验证以及属性简约过程[4]。

（1）数据质量检验：利用值分析方法去掉一些取值不正常、数值间无差异的指标。

（2）属性简约：属性简约可以帮助减少评级指标集中指标的数量，但仍然可以保持原数据的完整性。采用基于信息熵的属性简约算法，通过计算指标的重要度来筛选出重要指标。

2.4 模型求解

假设在属性简约处理后，所有指标都比较重要，没有剔除任何指标，则对字符型指标虚拟化处理以后，加上原来的指标，共有23个输入指标，即P=23。选取农村商业银行m组观测数据$(x_{i1},x_{i2}\cdots x_{ip},y_i)$，$i$=1,2，...，$m$为训练样本，代入公式（6），模型求解得到逻辑回归系数的估值$\hat{\beta}_0,\hat{\beta}_1\cdots\hat{\beta}_p$。

各指标的信用分数=逻辑回归系数×2/ln2+20，将各指标的信用得分加和，即得用户的信用总得分，根据分数划分信用等级。

当有新的农户数据集进入模型时，模型自动给出客户的信用分数，从而实现对客户信用等级的评价。

3 结语

文章采用统计学建模方法，首先简单介绍了逻辑回归模型的工作原理；然后选取客户年龄、受教育程度、收入来源、年收入、婚姻状况、贷款金额、还款方式、担保方式、贷款年限等评价指标作为指标体系，构建了基于逻辑回归的农村商业银行客户信用评级模型，对贷款客户的信用等级进行评价。该方法能够帮助银行有效区分违约客户，比较准确地量化客户违约风险，对于商业银行有效管控信贷风险具有非常重要的参考价值。

参考文献

[1] 廖绚,李兴绪.基于Logit模型的银行个人信贷风险管理评估[J].统计与决策,2008(21):50-52.
[2] 鄂英力.线性回归模型的一种有偏估计[D].渤海大学,2012.
[3] 赵慧.浅析个人信用评级模型[J].中国城市经济,2011(9):31.
[4] (美)RichardO.Duda,著.模式分类[M].北京:机械工业出版社,2003.

（上接254页）

增加，不共点只共线的开放面2，重复第二种结构第二（或三）开放面的性质。既不共点也不共线的开放面3，重复开放面1的属性特征，即重复开放面1的颜色，以此类推，在这种结构下无论数量多少则最多需要3种颜色。

同样新增加的只与开放面1共点的开放面，重复开放面1的属性特征，即重复开放面1的颜色，以此类推，在这种结构下无论数量多少则最多需要3种颜色。

（4）在属性范畴把两点一线的形状定义为节。封闭面与开放面共线的差异在于共节。因此需要增加一颜色区分开放面1与封闭面一的共节关系。在把开放面1和封闭面一组合为共同体作为一个新的起点后，开放面的演绎就与前面一、二、三的结构相同，如图5所示。

同样不考虑外部的开发面，从封闭面一出发，封闭面的变化也与前面一、二、三的结构相同。开放面与封闭面同步变化演绎，与组合体(1、一）非共线、非共点、非节的开放面或封闭面就是1或一属性特征重复的新起点。以此类推，在这种结构下无论数量多少则最多需要4种颜色。

3 结论

综上所述，可得出以下结论：在同一平面内，全息结构的属性关系只有3种，即共线、共节、共点，所以只需要4种颜色区分。并得出推论：把平面进行弯曲，只要不封闭，这个结论同样适用。

4 结语

道德经说：三十辐同一毂，当其无，有车之用也。分形结构为有，平面空间的反向结构为无，其用超过分形。四色地图问题中点、线、面的分形结构为有，反向结构区域的相邻与不相邻关系、封闭与开放关系为无，其用，成四色定理。

参考文献

[1] 张士庆,张号.四色问题的直观几何论证及单纯性地图四色定理[J].图学学报,2013(5):46-50.
[2] 肯尼思·法尔科内,著.分形几何——数学基础及应用[M].曾文曲,刘世耀,戴连贵,译.北京:人民邮电出版社,1991.