Statistical and Application 统计学与应用, 2013, 2, 51-54
http://dx.doi.org/10.12677/sa.2013.21006 Published Online March 2013 (http://www.hanspub.org/journal/sa.html)

A Study of Telecom Customer Loss Prediction Based on Generalized Linear Mixed Models*

Jun Wang[1], Yu Fei[1#], Jianxin Pan[2]

[1]School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming
[2]Yunnan University of Finance and Economics, Kunming
Email: dagaihaoxiang@gmail.com, #feiyukm@yahoo.com.cn

Received: Mar. 5th, 2013; revised: Mar. 14th, 2013; accepted: Mar. 20th, 2013

Abstract: With the increasingly fierce competition in the communication business and the growing importance of customer relationship management, how to improve the customer satisfaction and reduce the customer churn rate has became the main strategy to improve the competitiveness of telecom enterprises. Based on the summaries of former researches, this paper used the Generalized Linear Mixed Model to analyze the reasons of customer loss, find out the customer loss prediction model and provide several references for the telecom enterprises.

Keywords: Customer Loss; Generalized Linear Mixed Models (GLMM); Prediction

基于广义线性混合模型的电信客户流失预测研究*

王　珺[1]，费　宇[1#]，潘建新[2]

[1]云南财经大学统计与数学学院，昆明
[2]云南财经大学，昆明
Email: dagaihaoxiang@gmail.com, #feiyukm@yahoo.com.cn

收稿日期：2013 年 3 月 5 日；修回日期：2013 年 3 月 14 日；录用日期：2013 年 3 月 20 日

摘　要：随着通讯业务的竞争日趋激烈，客户关系管理的重要性日益突出，如何提高客户满意度、减少客户流失几率成为电信企业提高竞争力的重要策略。本文在总结国内外学者研究的基础上，采用广义线性混合模型分析客户流失原因，进行客户流失预测，为电信企业提供一定的参考。

关键词：客户流失；广义线性混合模型(GLMM)；预测

1. 引言

随着通讯业务的不断发展，我国的通讯企业逐渐进入缓慢增长期，在用户数量逐渐饱和的今日，越来越多的通讯企业意识到客户关系管理的重要性，分析客户流失原因、吸引潜在客户入网、增加现有客户满意度、减少客户流失几率、提高客户消费水平、充分占有市场，成了各个通讯企业面临的重要问题。据 Gartner 公司的统计调查数据显示，每减少 5%的客户流失，就可以增加 25%~85%的利润[1]。而发展一位新客户的成本是挽留一个老客户的 4~8 倍，向新客户进行推销的花费是向老客户推销花费的 6 倍以上[1]。由此可见，挽留老客户就等于是优化了营销渠道，降低了营销成本，提高了营业利润。因此，合理的利用客户关系管理、建立行之有效的客户流失预警模型，并根据分析的结果提出相应的对策，可以很好的提高企业的利润。针对这一研究问题，国内外学者进行了较

*基金项目：本研究是国家自然科学基金研究项目，项目名称：“线性混合模型和广义线性混合模型均值和方差协方差结构的同时拟合及其统计诊断”，项目批准号：11061036。
#通讯作者。

为广泛的研究。

邱义堂(2000)[2]以台湾某公司 GSM (全球移动通信系统)系统行动电话客户为研究对象，采用 C4.5 决策树方法并配合多专家决策分类方法建立预测，该模型能在 10%的客户群中预测到 50%以上的流失客户。Louis (2002)[3]引入决策树建立客户流失预测模型，与 Logistic Regression、判别分析等方法进行了对比分析。Cardeln (2003)[4]等人采用决策树 Tree-Net 算法对美国某公司的客户进行了流失预测，不仅获得了较高的命中率，而且获得了有效的识别客户流失的规则。Ultch (2002)[5]利用非监督学习方法中的 SOM (自组织映射)对客户进行流失分类，取得了较高的预测命中率和覆盖率。Au (2003)[6]等人认为流失预测模型不仅要预测客户流失或不流失，对于预测客户流失的可能性同样重要，因此采用 EL(进化学习)算法对客户流失进行了预测，与决策树 C4.5、SCS (A Simple Classifier System)和 GABL (GA Batch concept Learner)进行比较发现，EL 算法能得到更好的预测精度。Xu (2006)[7]等人针对客户流失数据的复杂性和噪声较大等特点，利用粗糙集和 BP-ANN (人工神经网络)对客户流失进行了预测。赵宇(2007)[8]等人针对美国 Duke 大学客户关系管理中心的调查数据，利用改进的 SVM (支持向量机)来预测未来可能流失的客户，取得了较高的整体准确率。Shao Jin-bo (2007)[9]等人为了平衡有数据抽样带来的预测偏差，引入了三种 AdaBoost 算法，并利用 SVM 方法建立了预测模型，得到了优化结果。

为了能进一步提高预测模型的精度和稳定性，本文拟将广义线性混合模型引入客户流失问题的研究中，试图通过设定随机效应，来识别出不同样本之间的异质性以及同一对象不同观测值之间相关性，使得到的客户流失预测模型具有较低的错判代价，从而为运营商的盈利提供理论依据。

2. 广义线性混合模型

广义线性混合模型(GLMM)[10]的基本条件是假定响应变量的分布是属于指数分布族的，并且在线性预测部分中引入了随机效应进行模型的拟合。在客户流失问题的研究中，响应变量 Y 通常表示的都是电信公司客户的在网状态，每一个观测值可以记为 y_i，用数值 0 或 1 来分别表示客户是离网或正常在网，属于二项分布。因此考虑采用对数形式的典则连接函数 $\eta = \log\left(\mu/(1-\mu)\right)$，将线性预测与 Y 的均值 μ 连接起来。从而，就可以得到关于客户流失问题研究的 GLMM 一般表达式：

$$\log it\left(y_i\right) = x_{ij}'\beta + \xi_i + \varepsilon_{ij} \tag{1}$$

其中，x_{ij} 是固定效应 β 对应的设计向量，ξ_i 是随机效应，ε_{ij} 是随机误差。

对于广义线性混合模型中 β 值的求解，可以考虑采用类似于广义线性模型(GLM)的极大似然估计方法得到结果，而当似然函数比较复杂时，就需要采用模型近似法、数值积分法或贝叶斯方法进行估计[10,11]。

3. 实证研究

3.1. 数据预处理

本文原始数据是来自云南省昆明市区以及所辖县区中国电信公司客户数据库中的 7241 位固定电话客户 2009 年 1 月~12 月每月的消费数据。其主要的指标有：所属区局、来显功能、彩铃功能、套餐名称、用户状态、1~12 月每月消费额。

通过初步的描述性统计，可以看出昆明市区四个分区的总样本数量为 4187 个，目前客户正常在网数量为 3524 个，则截止 2009 年市区客户正常在网的占比为 84.17%，离网占比为 15.83%。而县区局客户样本数是 3054 个，客户正常在网数量为 2206 个，所以，正常在网客户比例为 72.23%，离网率为 27.77%，这个数据高于市区客户的离网率。市区分公司标准资费客户离网率为 23.95%，而套餐资费型客户离网率仅为 2.85%；同样，县区局标准资费型客户发生离网的比例为 39.2%，套餐型客户离网比率为 5.71%。由此可以看出，无论是市区还是县区局，标准资费型客户发生离网流失情况的比率都远远大于套餐资费型客户。因此，对于电信公司而言，应着重分析标准资费型的客户。换言之，电信公司可以把客户流失预测分析的重点放在研究标准资费型客户上。针对此重点，本文就将套餐资费型客户排除在外，主要针对昆明市区和所辖县区局标准资费客户的消费情况进行分析和预测，得到的处理后的数据情况如表 1 所示。

3.2. 实证结果分析

在 GLMM 模型的框架下，再结合了昆明市电信

Table 1. Sample introduction
表 1. 样本情况介绍

样本类别	样本数目	所占比例
非正常客户	1405	30.6%
正常客户	3181	69.4%
总计	4586	100%

客户的数据特点，将拟合的模型定义为如下的形式：

$$\log it\left(y_{ij}\right)=x'_{ijk}\beta+D'_{il}\gamma+\xi'_{ij}\alpha+\varepsilon_{ij} \quad (2)$$

其中，x_{ijk}表示的是属于第 j 地区的第 i 个客户滞后 k 个月的消费情况，k 共取 4 种类型分别为 1、2、3、4，分别代表着滞后 12 月份 1、2、3、4 月的客户消费情况，即 8、9、10 和 11 月客户的消费额情况；D_{il}表示的是第 i 个客户第 l 种增值业务服务开通的情况，$l=1$ 代表的是来显功能的开通情况，$l=2$ 代表的是彩铃功能开通的情况；ξ_{ij}反应的是随机效应，即第 i 个客户是属于哪一个地区的，$j=1$ 表示的是客户属于市区，$j=0$ 表示的是客户属于县区局。

调用 R 软件中的 MASS 程序包，利用 glmmPLQ 算法，选用上述的指标对 12 月份客户的在网状态进行拟合。迭代 5 次后参数收敛，得到模型估计结果如下表 2 和表 3 所示。表 2 显示的是广义线性混合模型中固定效应参数的估计结果，从表中可以看出，客户流失主要受到彩铃开通情况、来显开通情况和第 8、9、10、11 月的消费额等固定效应的影响，而且消费额是越靠近预测月份则对客户状态指标的影响越大(回归模型以系数的绝对值大小来判断，越大说明该变量的影响越明显)，也就是说，最后四个月的月消费额越小，那么该名客户流失的可能性越大。从输出结果的 F 值及 F 值的伴随概率(Sig.)可以看出，每一个进入模型的固定效应都是显著的，拟合的效果还是非常不错的。表 3 显示的是随机效应的标准差。从两个表中可以看出来，模型拟合的效果是很好的。

为了检验模型拟合效果的优劣，本文采用回判分析的方法，将已知的数据代入预测模型，从而得到输出的预测结果，由此可以给出 GLMM 模型估计结果的正确率和错判率，如表 4 所示。

采用 GLM 方法回判结果见表 5；比较表 4 和表 5 可以看出，分类的结果中，采用 GLMM 回判不仅正常客户的判断正确率较 GLM 的结果有所上升，非正常客户的判断正确率也达到了 84%，从而使得总体的

Table 2. Parameter estimate of GLMM (1)
表 2. GLMM 参数估计(1)

固定效应	系数	t值	Sig.
常数项	6.916	39.621	0.000
来显功能	−1.342	97.188	0.000
彩铃功能	−1.889	43.838	0.000
8月消费额	0.009	14.403	0.000
9月消费额	−0.016	18.254	0.000
10月消费额	−0.025	32.645	0.000
11月消费额	−0.200	11.189	0.001

Table 3. Parameter estimate of GLMM (2)
表 3. GLMM 参数估计(2)

随机效应	StdDev.
常数项(region)	20.439
残差	13.66

Table 4. Back-contracting results based on GLMM
表 4. GLMM 回判结果分析

客户在网情况	数目	判断的客户在网情况	检验后数目	比率
正常客户	3181	“正常”客户	3104	97.6%
		“非正常”客户	77	2.4%
非正常客户	1405	“正常”客户	225	16%
		“非正常”客户	1180	84%
总计		准确率	93.4%	
		误判率	6.6%	

Table 5. Back-contracting results based on GLM
表 5. GLM 回判结果分析

客户在网情况	数目	判断后客户在网情况	检验后数目	比率
正常客户	3181	“正常”客户	3099	97.4%
		“非正常”客户	82	2.6%
非正常客户	1405	“正常”客户	270	19.2%
		“非正常”客户	1135	80.8%
总计		准确率	92.3%	
		误判率	7.7%	

预测准确率达到了93.4%，预测的结果比较理想。

然而，准确率的高低并不是比较模型优劣的唯一指标。本文采用较常出现在数据挖掘文献中的加权评价指标来对GLMM构建的预测模型和其他5种数据挖

Table 6. Comparison of the predictions among six methods
表 6. 各方法的预测结果比较

算法	第一类错误率	第二类错误率	精确率	覆盖率
GLM(Logistic)	0.1922	0.0258	0.9326	0.8078
GLMM	0.1601	0.0242	0.9387	0.8399
CHAID	0.2932	0.0340	0.9019	0.7068
CART	0.3402	0.0321	0.9009	0.6598
QUEST	0.4569	0.0119	0.9526	0.5431
多层感知机	0.3302	0.0380	0.8861	0.6698

掘算法得到的结果进行比较，如表 6 所示。

从表 6 中可以更加的直观的看出，在预测潜在的流失客户时，通过采用引入地区随机效应的 GLMM 的精确率、覆盖率和两类错误率分别为 0.9387、0.8399、0.1601 和 0.02421，除了第二类错误率略高于决策树算法中的 QUEST 算法，精确率低于 QUEST 算法外，其余的模型评价指标都高于其他算法所得到的结果。较高的精确率和覆盖率则显示出了 GLMM 预测结果的稳定性和可靠性；较低的两类错误率表示运营商可以适用较少的成本来预测到较多的潜在流失客户，从而使得运营商的盈利明显提升。

4. 结论

本文将 GLMM 引入客户流失预测模型的研究中，分析地区这一随机效应是否对客户流失问题存在着影响，地区这一指标包含两种类别：市区和县区局。基于客户流失问题的影响因素分析，从而得到对客户的在网状态进行判断的预测模型，目的是给电信公司提供相关建议。

本文的结论可以概括为以下方面：客户的月消费额会直接的反应出客户状态的变化趋势，尤其分析月份之前的 4 个月的消费情况，这 4 个月的消费额如果呈现出递减的趋势则表明该用户有很大的可能性会流失；标准资费型客户对于增值业务的选择是客户忠诚度的体现，选择的增值业务越多，则表明越满意该公司提供的服务，从而会长期的选择该公司的业务；不同的地区间，因为客户人群的分布不同也会导致客户的流失规则有所不同，采用统一的衡量标准会导致客户流失模型的预测准确率降低。

鉴于通讯行业的经营模式逐渐转变为“客户驱动”型，这就促使各企业更加积极地采取以客户为中心的营销维护策略，总结上述三项结论，可以将结论推广到国内的通讯行业并给出以下建议：

1) 从企业的营销角度来看，统计分析的结果指出固定套餐的用户的忠诚度普遍高于标准资费型的用户，并且，城市客户的流失率与县区局的客户存在着明显的差异。那么，企业可以根据客户数据中的这一特点，全方位、多样化的制定适合各种类型人群的套餐资费类型，从而提高套餐类用户的占比，保证企业拥有的固定客户数量在稳定中逐步上升。

2) 从企业的管理角度来看，管理层可以考虑将客户流失预测模型的结果与其他的一些相关指标结合起来，先将客户进行一定的细分后，再针对不同的流失客户制定有差异的挽回措施，这样就使得挽回的工作目的性提升，从而可以在保证客户挽回率的同时减少实施策略的挽回成本，以保证企业的长期发展。

参考文献 (References)

[1] 方红. 读者流失预警模型及其在公共图书情报机构中的应用[J]. 黑龙江科技信息, 2007, 2X(4): 103.

[2] 邱义堂. 通信资料库之资料挖掘: 客户流失预测之研究[D]. 国立中山大学资讯管理学系研究所, 2000.

[3] A. C. Louis. Data mining and causal modeling of customer. Telecommunication Systems, 2002, 21(2): 381-394.

[4] S. Cardelln, M. Golovnya and D. S. Inberg. Churn modeling for mobile telecommunications, 2003. http://www.salford-systems.com/doc/churnwinF08.pdf

[5] A. Ultch. Emergent self-organizing feature maps used for prediction and prevention of churn in mobile phone markets. Journal of Targeting Measurement and Analysis for Marketing, 2002, 10(4): 314-324.

[6] W. H. Au, K. C. C. Chen and X. Yao. A novel evolutionary data mining algorithm with applications to churn prediction. Evolutionary Computation, 2003, 7(6): 532-545.

[7] E. Xu, S. S. Liang and X. D. Gao. An algorithm for predicting customer churn via BP neural network based on rough set. Proceedings of Asia Pacific Conference on Services Computing. Washington DC: IEEE Computer Society, 2006: 47-50.

[8] 赵宇, 李兵, 李秀. 基于改进支持向量机的客户流失分析研究[J]. 计算机集成制造系统, 2007, 13(1): 202-207.

[9] J.-B. Shao, X. Li and W. H. Liu. The application of Ada-boost in customer churn prediction. Proceedings of International Conference on Service Systems and Service Management, 2007: 1-6.

[10] C. E. McCulloch, S. R. Searle and J. M. Neuhaus. Generalized, linear, and mixed models (2nd Edition). Wiley-Interscience, 2008.

[11] P. McCullagh, J. Nelder. Generalized linear models (2nd Edition). Boca Raton: Chapman and Hall, 1988.