DOI:10.16537/j.cnki.jynufe.2010.06.011

电信客户流失预警规则及其信度测定实证研究
——以云南电信为例

顾光同[1,2],王力宾[2],费 宇[2]

(1.浙江农林大学 理学院,杭州 311300;
2.云南财经大学 统计与数学学院,昆明 650221)

摘 要:以云南省昆明市区电信行业客户数据为样本,应用决策树算法分析流失客户特征,并得出电信客户流失的预警规则,计算出其预测准确率,再在显著的客户信息和预警规则的条件下,测定出所提取的规则发生的概率,即信度。实证研究表明:提取的预警规则准确率和信度都在95%以上,这为电信行业提供了有效的客户挽留决策分析依据。

关键词:电信客户流失;决策树;预警规则;准确率;信度

中图分类号:F224.7 文献标志码:A 文章编号:1674-4543(2010)06-0094-05

随着信息化时代的日新月异,电信业内运营商之间的竞争日趋激烈。Gartner公司调查数据表明,开发一个新客户的费用是维持一个老客户的4~5倍,[1]目前美国电信行业客户流失率为30%,欧洲约为35%~50%,这从客观上推动了客户流失预测系统的研究与应用。根据美国营销学者赖克海德和萨瑟的理论,一个公司如果将其客户流失率降低5%,利润就能增加25%~85%。[1]而电信企业从来不缺少客户的相关信息,比如年龄、套餐类型以及话费消费额等大量的数据,如何正确分析利用这些数据建立合理的客户流失预警模型,提高客户的忠诚度,防止客户流失,就成为每个电信运营商所亟待解决的问题。[2]

从研究趋势来看,国内近几年应用数据挖掘方法研究客户流失的文献比较多,其中有学者采用神经网络方法进行研究,其预测准确率为88.9%,但研究的样本数据是按流失客户个数与在网客户数量1∶1来选取的,[3]这意味着流失客户数量占整个样本的50%,这可能会导致所构建的预警模型过度拟合流失客户,从而使准确率大大提高。也有学者应用粗糙集理论进行客户流失预警,其预测准确率为88.3%。有的采用支持向量机(SVM)方法,可识别60%的客户。总体来看,各文献所采用的研究方法,其预测误判率都至少在10%以上,预测结果不太理想。众所周知,企业往往获得一位新客户的费用是比较高的,但是挽留一个客户有时仅仅只需一个电话或者一个策略。[4]因此,如何分析利用流失数据或经验,提取流失客户的特征规则,提高对即将流失的客户的预警准确率,最终为电信行业采取合理、科学的挽留措施,成功地挽留客户提供了有力的依据。

基于云南省昆明市电信行业客户信息资料,将决策树算法和Logistic回归法结合,分析电信客户特征,力求构建一个高效、科学的客户流失预警模型,为电信行业提供相应的有效客户挽留策略。

收稿日期:2010-07-14

基金项目:云南省教育厅科研基金项目"电信客户流失的统计建模与预测"(09J0011)

作者简介:顾光同(1977-),男,云南宣威人,浙江农林大学理学院讲师,云南财经大学统计与数学学院硕士研究生,研究方向为统计理论法与应用、经济统计;王力宾(1958-),男,山西安泽人,云南财经大学教授,博士,硕士生导师,研究方向为国民经济统计学;费宇(1968-),男,云南会泽人,云南财经大学统计与数学学院教授,博士,硕士生导师,研究方向为统计理论与方法、计量经济理论和方法。

一、研究方法及思路

采用 Magidson提出的决策树非线性多元技术 CHAID(CHi-squared Automatic Interaction Detector)算法,[5] 对企业现有的大量客户信息进行客户特征分析, 计算出客户特征的信息熵, 将电信样本数据按树状结构分成若干分支形成决策树, 每个分支包含客户特征包的类别归属共性, 从每个分支中提取有用的客户信息特征, 得到客户流失的主要 k条预警规则, 这里记第 i条规则为事件集 C_i($i=1, 2, \cdots, k$)。在电信客户决策树的生成过程中, 其输入为训练样本数据集, 决策树是其最终的输出结果, 该树的每一个节点对应着客户特征包进行分类的一个决策属性, 其分支对应着客户特征包按该属性进一步划分取值特征, 叶子节点代表着电信客户的各个类或类的分布情况。为了避免因数据集规模小或噪音大, 造成流失客户过度拟合的情况, 首先按流失客户占在网客户 3∶7的比例选取样本, 然后随机选取 50%的样本数据作为训练集, 剩下 50%的为测试集。最终, 提出合理、科学的预警规则。

为进一步研究预警规则的可信度, 将基于决策树算法提取的客户流失预警规则 C_i($i=1, 2, \cdots, k$)引入二元 Logistic回归模型, 那么在事件 $\{X_1=x_1, X_2=x_2, \cdots, X_p=x_p\}$和 C_i同时发生条件下的概率预测模型则为:

$$\pi_i = P(Y=1 \mid \{X_1=x_1, X_2=x_2, \cdots, X_p=x_p\} \cap C_i) = \frac{1}{1+e^{-a-b_1x_1|C_i-\cdots-b_nx_p|C_i}} \quad (1)$$

其中, $a>0$, $b_i\geq 0$($i=1, 2, \cdots, p$), $X_1, X_2, \cdots, X_p$表示电信客户信息变量, $Y=1$表示客户流失事情发生, π_i表示第 i条预警规则发生的概率, 文中称其为客户流失预警规则的信度。

下面以云南省电信企业提供的历史数据为分析对象, 应用决策树算法分析流失客户和在网客户的信息特征, 得到客户流失的信息规则, 并给出其预警能力的准确率, 也就是该规则的预警精度, 根据该规则, 分析并判断现有客户是否会流失, 然后, 再应用 Logistic回归方法测定出该规则发生的概率, 也就是预警规则的信度。

二、数据准备

1. 样本选择与数据来源

实证数据来源于昆明市电信公司共 2565名固定电话客户 2008年 1~7月的消费额样本数据。其主要指标如下: 客户年龄、性别、资费类型、客户状态、1~7月客户话费消费额。其中客户资费套餐类型包含三类, 分别是标准资费、套餐类资费以及其他类资费类型。客户当期状态指标包含正常或非正常, 其中正常即是客户在网, 不正常即为客户流失。对该数据预处理原则如下: 一是考虑到缺失值或异常值对统计分析结果的不利影响, 直接剔除缺失的数据和消费额波动大的数据。二是由于客户套餐资费类型繁多, 剔除数据中套餐类型比重小于 5%的客户数据。三是为提出科学、合理的客户挽留策略, 实证分析时, 过滤掉不显著相关的信息。按照上述原则预处理方法得到标准资费套餐和非标准资费套餐 (这两类套餐类型占总套餐类型的 97.9%)下共 1711个客户样本数据, 其中流失客户与在网客户的比例为 3∶7, 即在实证分析的样本数据中流失客户占 30%, 这和目前电信行业的客户流失的真实数值相近, 见表 1所示。

表 1　　样本电信公司客户状态变量频数分析结果

	频　率	有效百分比
流失客户 (1)	530	31.0
在网客户 (0)	1181	69.0
合计	1711	100.0

注: 根据昆明市电信公司提供的电信数据, 将客户状态进行数据预处理, 分为流失客户和在网客户, 分别用 1和 0表示。

本文的数据预处理、计算[6]主要使用统计软件 PASW18.0以及数据挖掘软件 PASW Modeler13.0完成。

2. 研究指标的选取

基于上述预处理后的数据，用变量 X_i 表示第 i 月（$i=1, 2, \cdots, 7$）的话费消费额，虚拟变量 $D=\begin{cases}1 & \text{表示套餐类型}\\ 0 & \text{非标准资费套餐类}\end{cases}$ 表示套餐类型，X_8 表示客户年龄，对于客户性别等其他指标，由于和客户流失的关系不显著，故不选取。因此，电信客户流失预警规则的输入变量为 X_1，X_2，…，X_7，X_8，D，输出变量即被解释变量用二值变量 $Y=\begin{cases}1 & \text{客户流失}\\ 0 & \text{客户在网}\end{cases}$ 表示。

三、实证结果与分析

根据提出的研究方法以及研究思路，得到下面基于预处理后的样本数据。首先，用决策树算法得出客户流失预警规则，并计算出其精度，然后，应用本文提出的概率预测模型，测定其信度。

1. 预警规则的得出及其预测准确率

选用决策树算法中目前流行的穷举 CHAID算法来进行客户特征分析，得出客户流失预警规则，随机选取 1∶1的训练集和测试集，分别作为训练样本和测试样本。借用数据挖掘软件 PASW Modeler13.0，选用 Bonferron方法调整各节点的显著性值，穷举 CHAID算法的收敛阈值设定为0.001，迭代次数设为 500次，再选用较稳健的似然比统计量为分支的显著性检验统计量，得到基于测试集输出的电信客户决策树，见图 1所示。其指标的重要性依次为 D，X_7，X_6，即电信客户流失预警中最重要的信息是客户的资费套餐类型，其次就是客户的近两个月的话费消费额度信息。

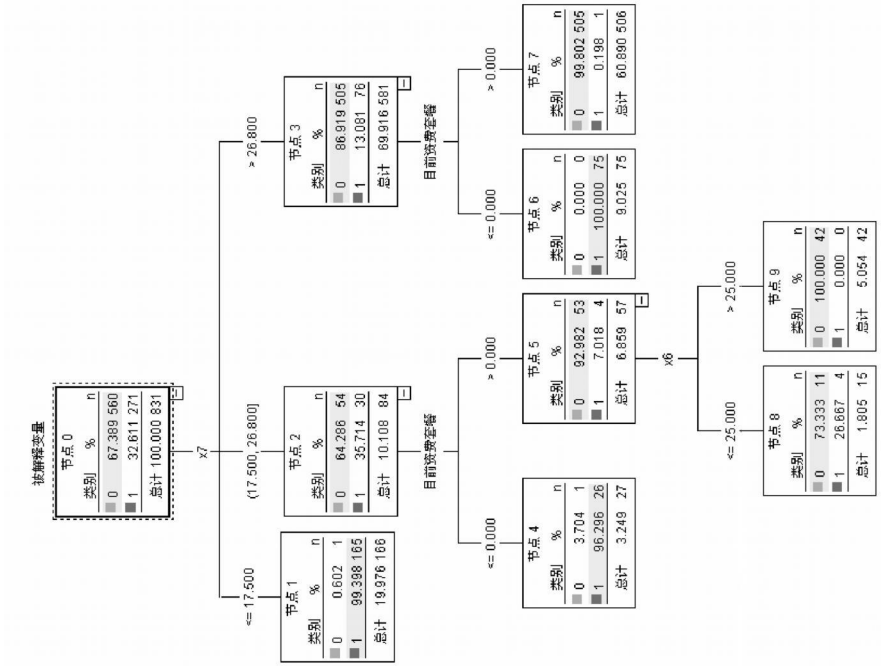

图 1 电信客户穷举 CHAID决策树

根据图 1，得出主要的客户流失预警规则及其预测准确率。

（1）规则 C_1：若某客户前一个月的话费总额小于等于 17.5元，则认为该客户有 99.4%的可能性会流失。

（2）规则 C_2：若某客户前一个月的话费总额大于 17.5元且小于等于 26.8元，目前资费套餐为标准资费类型，而且再前一个月的话费总额大于 25.0元，则认为该客户不会流失，即该客户在网，反之，

则认为该客户一定会流失。

(3)规则 C_3: 若客户前一个月的话费总额大于 17.5元且小于等于 26.8元, 目前资费套餐为非标准资费类型, 则认为该客户有 96.3%的可能性会流失。

(4)规则 C_4: 若客户前一个月的话费总额若大于 26.8元, 目前资费套餐为非标准资费类型, 则认为该客户一定会流失。

2. 客户流失预警规则的信度测定

分析上述四个规则的有效性, 最好的方法就是预测出该规则发生的概率大小, 即客户流失预警规则的信度。利用统计软件 PASW18.0 首先, 应用二元 Logistic回归方法, 以 Y为因变量, X_2, X_5, …, X_8, D为协变量, 其中 D还是一个分类变量, 用逐步向前似然比进行变量筛选, 算法迭代收敛阈值设定为 0.001。计算得初始似然比统计量为 $-2\ln(LR)=2117.906$, 程序共进行了 8次迭代, 最终, $-2\ln(LR)=151.867$, 该值的快速下降, 说明模型拟合得更加有效。比较稳健的 Nagelkerke拟合度 R^2 值为 0.962, 进一步说明模型拟合较好。参数估计结果见表 2。

表 2　　二元 Logistic回归分析结果

变量名称	系数	标准误	Wald统计量	自由度 df	伴随概率 Sig.	Exp(B) 置信界限
目前资费套餐 D	21.770***	1.989	119.811	1	0.000	0.000
年龄 x_8	−0.046**	0.020	5.207	1	0.022	1.047
消费额 x_2	0.027***	0.008	9.806	1	0.002	0.974
消费额 x_5	−0.118***	0.018	44.906	1	0.000	1.125
消费额 x_6	−0.132***	0.020	44.325	1	0.000	1.141
消费额 x_7	−0.173***	0.016	113.900	1	0.000	1.189
常量	9.804***	1.424	47.383	1	0.000	0.000

注: “**”、“***”分别表示在 5%、1%的显著性水平下通过了统计检验, 该结果通过 PASW18.0分析电信公司样本数据整理而得。

由此可得到客户流失的 Logistic概率预测模型为:

$$P(\text{客户流失}\mid X_2, X_5, X_6, X_7, X_8, D)=\frac{1}{1+e^{9.804-0.027X_2+0.118X_5+0.132X_6+0.173X_7+0.046X_8-21.770D}} \quad (2)$$

由模型(2)可知, 客户流失主要受第 2、5、6、7月的消费额以及年龄和资费套餐影响比较大, 尤其该客户近 3个月的消费额越小, 说明该客户流失的可能性越大。对估计系数取绝对值, 根据绝对值的大小可知(2)中主要变量的重要程度与决策树得到的结果一致。

结合模型(1)和(2), 可以进一步研究基于决策树得到的客户流失预警规则的有效性, 也就是该规则发生的概率, 即信度。

(1)规则 C_1 的信度: $\pi_1=P(Y=1\mid\{X_2, X_5, X_6, X_7, X_8, D\}\cap C_1)=0.9790$

(2)规则 C_2 的信度: $\pi_2=P(Y=1\mid\{X_2, X_5, X_6, X_7, X_8, D\}\cap C_2)=0.9998$

(3)规则 C_3 的信度: $\pi_3=P(Y=1\mid X_2, X_5, X_6, X_7, X_8, D\}\cap C_3)=0.9532$

(4)规则 C_4 的信度: $\pi_4=P(Y=1\mid\{X_2, X_5, X_6, X_7, X_8, D\}\cap C_4)=0.9862$

由此可知, 决策树算法得出的流失客户预警规则其信度都在 95%以上, 因此, 进一步说明提取的 4个预警规则适合为电信行业进行客户挽留提供决策分析依据。

四、结论与建议

目前大部分省级和地市级移动公司都已经建立了各自的业务系统, 业务信息包括客户基本信息、大客户信息、收益信息、市场竞争、服务质量、营销管理、语音业务、新业务及数据业务、合作服务信息等资料。采用合理的统计分析对客户流失进行预警, 能较好地帮助电信企业提升管理水平, 从而提高企业的利润, 降低成本。以云南电信为例的客户流失统计分析结果, 可归纳出客户流失的共有特征,

并提出相应的挽留政策,目的是降低客户流失率。[7] 下面提出几点结论和建议。

第一,电信行业在提取有用信息有效分析客户特征时,应注意:一是样本数据的选取,尤其客户流失所占的比重应该跟实际值相符。二是应该对缺失值以及异常值的处理和不显著信息进行过滤。三是选择合适的模型或算法,尤其合理设定算法的迭代方法、迭代次数、迭代精度以及变量的检验统计量以及进入或剔除标准。

第二,得出的客户流失预警规则或者建立预警模型的目的都是为了挽留客户,防止流失,本文针对性地提出几点挽留措施:一是根据电信企业提供的客户信息来看,套餐模式中流失的客户仅占2.24%,因此,应积极向客户推广套餐模式。话费方面,只要客户的近两三个月话费额始终在某个话费范围,那么,该客户一般不会流失。为了挽留住现有客户,运营商应该对客户进行差异化和特色化服务。二是从得出的预警规则来看,最有价值的信息是客户近两个月的话费消费额以及套餐资费类型。因此,电信企业不但应该在客户资费套餐的方式上下功夫,还应建立一个客户话费消费额的跟踪平台。

第三,从企业的营销角度来讲,企业应该根据决定客户流失的重要因素改变传统的营销观念,在继承传统营销中的有效策略的同时应弥补现有的不足,提高客户忠诚度,如将营销中工作人员的“一次性交易的顾客”价值观念转变为“终身”价值观念,让渡顾客价值,考虑客户成本和利益,与顾客采取多元化合作策略形成多次交易或者长期交易,也可以跟某些客户形成买卖合作乃至战略合作关系,同时,企业还应该分析客户的话费等生命周期。

第四,从管理角度来讲,管理层可以将所建的客户流失预测模型与客户的其他属性特征结合起来,对那些价值高流失倾向大的客户优先采取相应的挽留措施,以保证优质客户的持有率。必要时,可以先进行客户细分进行建模,这样会使问题的解决更有针对性。因此,客户细分后的分组模型的建立值得进一步研究。

参考文献:

[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,(4):103
[2] 罗布·马蒂森.电信业客户流失管理——电信管理精选译丛[M].北京:人民邮电出版社,2006:4—26.
[3] 夏国恩,金炜东.基于支持向量机的客户流失预测模型[J].系统工程理论与实践,2008,(1):76—77.
[4] 严伟.如何防范客户流失[J].企业管理,2003,(6):52—54
[5] 孙红,朱雷,刘毅婷.决策树在寿险企业客户流失分析中的应用[J].现代商业,2008,(20).
[6] 王力宾,顾光同.多元统计分析:模型、案例及SPSS应用[M].北京:经济科学出版社,2010:66—196
[7] 曹亚东,磨莉萍.客户保持—电信企业的新课题[N],人民邮电报,2003—08—11.

责任编辑、校对:李品秀

An Empirical Study on the Early Warning Rules of Telecom Customer Churn and Its Reliability Determination
— Taking Yunnan's Telecom Industry as an Example

GU Guang—tong[2], WANG Li—bin[2], FEI Yu[2]
(1. School of Sciences, Zhejiang Agriculture and Forestry University, Hang zhou 311300, China;
2. School of Statistics and Maths, Yunnan University of Finance and Economics, Kunming 650221, China)

Abstract: By taking the customer data of telecom industry in Kunming city of Yunnan Province as samples, the authors use decision tree algorithm to analyze the characteristics of loss customers, find out the early warning rules of telecom customer churn, and calculate the predicting accuracy. Besides, based on the significant customer information and the early warning rules, the authors measure the probability that the rules may occur, that is, the reliability. The empirical study shows that the accuracy and the reliability of the early warning rules are above 95%, which provides effective reference to customer retention analysis for telecom industry.

Key words: Telecom Customer Churn; Decision Tree; Early Warning Rules; Accuracy; Reliability