Doi: 10.3969/j.issn.1003-5060.2010.11.024

基于 Cox 模型的移动通信行业中低端客户流失预测研究

邓森文, 马溪骏
(合肥工业大学 管理学院, 安徽 合肥 230009)

摘 要: 文章将 Cox 模型运用于客户流失预测研究中, 通过对中国移动通信某分公司提供的历史数据的研究, 计算出每个客户的生存概率, 并按照生存概率从小到大排序, 等分为 10 组, 实际流失的客户基本上都落在第 1 组中, 覆盖率达到 89%以上, 运营商只对原来客户的 10%进行维护与挽留, 大大减少了工作量和挽留成本。该方法用于移动通信行业的客户流失预测中有很好的效果。

关键词: 客户流失; Cox 模型; 流失预测
中图分类号: TP311.13 文献标志码: A 文章编号: 1003-5060(2010)11-1698-04

Research on mid-and-low customer churn prediction based on Cox model in mobile telecommunication industry

DENG Sen-wen, MA Xi-jun
(School of Management, Hefei University of Technology, Hefei 230009, China)

Abstract: The Cox model is applied in customer churn prediction in this paper. By studying the historical data provided by a subsidiary of China Mobile Telecommunication, the survival probability of each client is calculated, and then it is sorted into ten groups according to the data sequence from small to large. The actual loss of customers basically comes into the first group and the coverage rate can be up to 89% or more. Operators only need to maintain 10% of the original customers, thus greatly reducing the workload and maintenance costs. This method has a good effect on customer churn prediction for the mobile telecommunication industry.

Key words: customer churn; Cox model; churn prediction

近年来, 客户流失已成为全球电信企业面临的一个普遍性问题。目前对电信业客户流失预测问题的研究十分广泛, 运用最广泛的是决策树算法[1]。决策树算法建模简单、分类准确率高, 而且能导出简明易懂的诸如 If-Then 形式的分类规则, 但也有一定的缺点, 此外, 很多专家对 Logistic 回归、人工神经网络和贝叶斯网络等方法[2-5]也进行了研究, 但是整个神经网络的分析过程是一个不透明的"黑盒子", 无法展现可读的模型, 每阶段的加权与转换亦不明确显示, 所以神经网络大多数都用于处理高度非线性且变量有相当程度交互效应的数据。

利用 Cox 生存分析建模算法预测客户流失问题有以下优点: ① 既考虑危险(流失)事件"发生"或"不发生"的结局, 也充分利用生存时间的信息; ② 能够处理删失数据。在生存分析中, 观测期截止时尚未流失的客户可以作为删失样本进入模型, 从而提高了模型的实效性, 且有利于模型的

收稿日期: 2009-10-16
基金项目: 国家自然科学基金资助项目(70631003); 高等学校博士学科点专项科研基金资助项目(200803590007)
作者简介: 邓森文(1982—), 男, 广西桂林人, 合肥工业大学硕士生;
马溪骏(1952—), 女, 安徽六安人, 合肥工业大学研究员, 硕士生导师.

实时更新。鉴于此, 本文利用 Cox 生存分析建模算法, 把已有数据分为训练样本和测试样本。通过训练样本, 利用偏最大似然参数估计方法计算出模型中每个属性的系数的估计值, 建立模型, 然后计算测试样本中每个客户的生存概率, 按生存概率从小到大进行排序, 等分为 10 组, 计算第 1 组包含流失的客户百分比, 这个比值越高, 模型的效果就越好。这样建立的基于电信行业客户流失的预测模型, 可以大大提高预测准确率, 为电信企业的客户保持和客户挽留提供有力的决策支持。

1　Cox 模型简介

1.1　生存时间函数

生存时间测量某事件出现的时间, 通常用生存函数、概率密度函数和危险率函数来描述。三者在数学上是等价的, 得出其中 1 个, 就可以推导出另 2 个。

生存函数(survival function), 又称累计生存率, 是指个体生存时间大于 t 的概率, 即

$$S(t) = P(T \geqslant t) = 1 - F(t) \quad (1)$$

其中, $F(t)$ 指个体的生存时间 T 的分布函数。

概率密度函数(probability density function), 又称作密度函数, 该函数的图形为密度曲线, 在任何时间区间内死亡的比例和死亡出现的机会峰值均可从密度曲线找出, 函数表达式为:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(\text{个体在区间}(t, t+\Delta t)\text{中死亡})}{\Delta t} \quad (2)$$

危险率函数(hazard function), 又称为风险函数、瞬间死亡率、死亡强度、条件死亡率、危险率等, 危险率函数是生存分析最基本的函数, 即

$$h(t) = \lim_{\Delta t \to 0}(P(\text{年龄是 } t \text{ 的个体在}(t, t+\Delta t)\text{ 中死亡}))/\Delta t \quad (3)$$

对于危险率函数, 有:

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} = -\frac{\mathrm{d}\ln S(t)}{\mathrm{d}t} \quad (4)$$

即

$$S(t) = \exp\left[-\int_0^t h(u)\,\mathrm{d}u\right] \quad (5)$$

1.2　*Cox* 模型

Cox 模型[6-8] 在表达形式上与参数模型相似, 但对各参数进行估计时又不依赖特定分布的假设, 所以又称为半参数回归模型。当生存时间是连续分布且预后变量间相互作用可被忽视时, 危险率函数 $h(t)$ 为:

$$h(t) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) \quad (6)$$

其中, h_0 为基准的生存分布的危险率函数; β 为回归系数; x 为预后变量, 即为协变量。由于 Cox 模型的假设, 每个预后变量的危险率在时间上正比于基准危险率 h_0, 从而无需计算 h_0, 使用起来非常方便。这时, 相应的生存函数为:

$$S(t; X) = S_0(t)^{\exp(\beta X)} \quad (7)$$

其中, $S_0(t)$ 为 t 时刻的基准生存函数。

在时间 t 和协变量 X 的作用下, 个体风险函数相对于基准风险函数之比与时间无关, 不随时间 t 的变化而变化; 而基准风险函数 $h_0(t)$ 只与时间 t 有关, 不受 X 的影响。Cox 模型不仅可以分析各协变量对生存时间的影响, 而且对基准风险分布不作任何要求, 就可以处理时变协变量。

2　基于 *Cox* 模型建立预测模型

2.1　数据准备

本文利用中国移动通信行业某分公司的客户进行实证研究。为了避免学生毕业和民工返乡造成的无法挽留的客户流失, 本文采集了 2007 年 1 月到 2007 年 6 月的数据, 其中 1 ~ 4 月为数据观测期, 该期间的客户基本资料、通话记录、账单等转化为属性后作为模型的输入变量, 6 月份的流失数据作为模型的输出。为了更好地刻画客户的消费行为, 本文引入月均话费、月均短信费用等[9] 一些衍生的属性。

根据本文算法, 生存分析中变量主要分为 3 类: 生存时间 T、删失变量 C 及表示相关因素的协变量 X。其中生存时间 T 定义为客户从开户到流失或者删失的时间, 以月为单位。由于电信行业客户不像其它行业的产品有固定的截止日期, 只要到观测期结束还没有流失的样本都是删失样本。因此, 如果客户到 2007 年 6 月份还没有流失, 则定义为删失样本, $C = 0$, 否则, 对于已经观测到流失的客户, $C = 1$; 影响流失行为的协变量总共为 12 个, 定义为 $x_1 \sim x_{12}$。

经过数据清洗与处理, 从数据库中得到 159 177 个资料完整的客户样本, 其中流失客户数为 14 776 个, 流失客户占比为 9.28%。然后按照 1∶1 左右的比例划分训练样本集和验证样本集, 其中, 训练样本集包含 72 843 个客户样本, 流失客户数为 7 482 个; 验证样本集包含 86 334 个客户样本, 流失客户数为 7 294 个。

2.2　建模属性处理

数据集中的属性较多, 其中有些属性可能与客户流失的相关性较大, 而有些可能与客户流失无关, 而且有些属性之间存在强相关关系, 即冗余属

性,因此要对属性进行约简。本文使用 Pearson 相关系数检验、Kendall′s tau-b 及 Spearman 秩次相关系数[3]来检验 2 个变量之间的相关性,以此来消除冗余。

Pearson 检验 2 个变量之间是否存在线性相关关系,如果变量 X 与变量 Y 呈完全正线性相关关系,则该系数等于 1;如果变量 X 与变量 Y 呈完全负线性相关关系,则该系数等于 -1;如果变量 X 与变量 Y 没有任何线性相关关系,则该系数等于 0,用公式表示为:

$$\rho_{xy}=\frac{\mathrm{Cov}(x,y)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}}=\frac{E((x-E(x))(y-E(y))}{\sqrt{E(x-E(x))^2E(y-E(y))^2}} \tag{8}$$

与 Pearson 不同,Spearman 只检验变量间的单调关系,而不强调线性相关,如果该系数等于 1,说明变量 Y 是变量 X 的完全增函数,但并不表示变量 X 和变量 Y 之间有任何线性相关关系,用公式表示为:

$$\tau=2(n_x-n_y)/[n(n-1)] \tag{9}$$

其中,对于(x_1,y_1)和(x_2,y_2),定义

$$\mathrm{sgn}\ x=\begin{cases}-1, & x<0;\\ 0, & x=0;\\ 1, & x>0\end{cases} \tag{10}$$

则 $n_x=\mathrm{sgn}(x_2-x_1)=\mathrm{sgn}(y_2-y_1)$;

$n_y=\mathrm{sgn}(x_2-x_1)=-\mathrm{sgn}(y_2-y_1)$。

Kendall′s tau-b 是一种对 2 个有序变量或 2 个秩变量间的关系程度的测度,因此也属于一种非参测度,其表达式为:

$$\theta=\frac{\sum_i((R_i-R)(S_i-S))}{\sqrt{\sum_i(R_i-R)^2\sum_i(S_i-S)^2}} \tag{11}$$

根据以上 3 种检验方法,计算各协变量与客户流失相关性的检验结果见表 1 所列。表 1 中 x_1 ~ x_{12}对应的属性分别为:x_1,年龄;x_2,性别;x_3,区域;x_4 是否有联系方式;x_5,是否本地身份证;x_6,总欠费次数;x_7,呼叫次数;x_8,月均短信费用;x_9,月均总费用;x_{10},平均开通业务数;x_{11},是否漫游;x_{12},信用度。

从表 1 可知,3 种检验方法的结果基本一致。在 0.05 的显著性水平下,年龄、是否有联系方式、是否本地身份证、呼叫次数、月均短信费用、月均总费用、平均开通业务数等 7 个属性与客户流失有显著的负相关关系;总欠费次数与客户流失有显著的正相关关系;性格、是否漫游、信用度和区域等 4 个属性与客户流失的关系不显著。取线性关系最强的 8 个变量作为模型的最终协变量。通过基于累计风险函数图示法来检验[10],以上筛选出来的 8 个协变量都满足 PH 假定。

表 1 各协变量与客户流失相关性检验结果

协变量	Pearson 流 失	Kendall′s tau_b 流 失	Spearman′s rho 流 失
x_1	-0.022 <0.000 1	-0.020 <0.000 1	-0.024 <0.000 1
x_2	0.000 0.924	0.000 0.924	0.000 0.924
x_3	0.064 0.728	0.070 0.728	0.081 0.728
x_4	-0.671 <0.000 1	-0.671 <0.000 1	-0.671 <0.000 1
x_5	-0.051 <0.000 1	-0.051 <0.000 1	-0.051 <0.000 1
x_6	0.313 <0.000 1	0.313 <0.000 1	0.313 <0.000 1
x_7	-0.052 <0.000 1	-0.052 <0.000 1	-0.052 <0.000 1
x_8	-0.363 <0.000 1	-0.363 <0.000 1	-0.363 <0.000 1
x_9	-0.667 <0.000 1	-0.667 <0.000 1	-0.667 <0.000 1
x_{10}	-0.640 <0.000 1	-0.640 <0.000 1	-0.640 <0.000 1
x_{11}	0.000 0.236	0.000 0.236	0.000 0.236
x_{12}	-0.533 0.352	-0.533 0.352	-0.533 0.352

2.3 模型的建立

根据(7)式,利用偏最大似然参数估计方法(Partial Maximum Likelihood)估计系数 β。本文使用 SPSS[6]统计软件,利用上述筛选出来的 8 个变量对训练样本进行拟合。参数估计结果见表 2 所列。

表 2 模型中 8 个协变量的估计结果

属 性	B	SE	Wald	df	Sig
年 龄	-0.012	0.001	96.768	1	<0.000 1
是否有联系方式	-0.771	0.040	365.587	1	<0.000 1
是否本地身份证	-0.703	0.024	828.860	1	<0.000 1
总欠费次数	0.992	0.046	473.818	1	<0.000 1
月均呼叫次数	0.868	0.155	31.249	1	<0.000 1
月均短信费用	-0.729	0.055	173.919	1	<0.000 1
月均总费用	0.332	0.057	34.063	1	<0.000 1
月均开通业务数	-1.628	0.073	495.734	1	<0.000 1

从表 2 可以看出,8 个预测协变量都在 0.01

置信水平上显著; 自由度为 1; 回归系数标准误差都很小, 说明用这些属性来预测客户流失的可靠性是比较大的。

年龄、是否有联系方式、是否本地身份证、月均短信费用、月均开通业务数的系数均为负值, 表明与客户流失负相关; 欠费次数、月均呼叫次数、月均总费用的系数均为正值, 表明与客户流失正相关。

2 4　模型的评价

在建立模型后, 将测试样本的预测变量值带入模型, 根据生存概率公式计算可以得到每个客户在 2007 年 6 月份的生存概率, 然后按客户生存函数排序, 将样本客户按照其在特定时点的预测生存概率从小到大排序, 等分为若干组, 比较各组中在预测的时间点之前流失的客户数量, 如果模型预测能力足够强, 该时间点越靠前, 即预测生存概率值越小的组中实际流失客户数应该越多。

本文按预测的客户流失率大小等分为 10 组, 然后计算每组中客户流失数, 如果模型的预测效果很好, 则每组的客户流失数应该递减, 且区别较大。而前面几组中包含的实际流失的客户占流失客户总数的百分比越高, 说明模型预测能力越好, 也越实用。按照上述方法进行分类, 结果见表 3 所列。

表 3　按概率排序分类结果

组　数	1	2	3	4	5	6	7	8	9	10
流失量/ 个	6 517	384	226	116	28	14	9	0	0	0
比率/ %	89 35	5 26	3 10	1 59	0 38	0 19	0 12	0	0	0

从表 3 可以看出, 改变预测模型确实能够将客户流失率按大小有效地区别开, 在选取的 2007 年 6 月份这个时间点上, 生存函数预测值最小的一组包含 89. 35% 的流失客户, 前 2 组基本上能涵盖 94 61% 以上的流失客户, 并且第 1 组总共有 8 633 个客户, 流失客户数占 75. 49%。因此, 利用 Cox 模型, 用预测生存概率最小的前 10% 的客户就能包含 89% 以上的实际流失客户, 这样, 只要集中资源对这 10% 的客户采取有效的针对性维护措施, 就有可能挽留住绝大部分可能流失的客户, 从而提高资源利用率, 最大程度降低客户流失率。

实证结果表明, 本文所使用的客户流失预测模型的预测效果是令人满意的。

3　结束语

本文基于 Cox 生存分析方法的客户流失预测模型在实际应用中还需要及时更新, 因为模型的训练是基于一个时间段内的数据进行的。该模型往往只代表了一段时间内用户的消费习惯和消费结构, 因此用模型预测时, 其时效性是明显的。当市场环境、用户的行为发生改变时, 模型也需要及时更新, 使用新的数据进行训练, 不断进行修正和完善以保证其有效性, 随着训练样本的增大, 本模型在预测命中率和预测覆盖率方面还有待于进一步提高。

[参　考　文　献]

[1] 盛昭瀚, 柳炳祥. 客户流失危机分析的决策树方法[J]. 管理科学学报, 2005, 8(2): 20－25.

[2] 王　雷, 陈松林, 顾学道. 客户流失预警模型及其在电信企业的应用[J]. 电信科学, 2006, 22(9): 47－51.

[3] 夏国恩, 陈　云, 金炜东. 电信企业客户流失预测模型[J]. 统计与决策, 2006, (20): 163－164.

[4] 贾　琳, 李　明. 基于数据挖掘的电信客户流失模型的建立与实现[J]. 计算机工程与应用, 2004, 40(4): 185－187.

[5] Mozer M C. Predicting subscriber dissatisfaction and improving retention in the wierless telecommunications industry [C]//IEEE Trans on Neural Networks, 2000, 11(3): 690－699.

[6] 卢纹岱. SPSS for Windows 统计分析 [M]. 第 3 版. 北京: 电子工业出版社, 2006: 571－578.

[7] 余红梅. Cox 比例危险回归模型诊断及预测有关问题的研究[D]. 西安: 第四军医大学卫生统计学研究室, 1998.

[8] Cox D R. Regression models and life-tables (with discussion)[J]. Journal of the Royal Statistical Society: Series B, 1972, 74: 187－220.

[9] 刘绍清, 黄章树. 生存分析在电信增值服务行业客户流失分析中的应用[J]. 广州大学学报: 自然科学版, 2006, 5(6): 33－36.

[10] 余红梅, 何大为. 检查 Cox 模型比例风险假定的几种图示法[J]. 中国卫生统计, 2000, 17(4): 215－218.

(责任编辑　闫杏丽)