客户流失预测模型构建

——基于生存分析视角

叶敏 季国民 张希君

(福州大学至诚学院 经济管理系,福建 福州 350004)

摘 要:客户是企业生存之本。随着企业间竞争的加剧,企业对客户的争夺也越来越激烈。有研究表明,获取 1 位新客户的成本大约是维护 1 位老客户成本的 5 到 6 倍,因此,如何维护现有的客户关系,防止客户流失,对企业生产尤为重要。通过对客户流失相关理论研究,分析客户流失的动因,寻找客户的特征,可以构建一个企业客户流失预测模型。该模型通过以客户与企业沟通的频繁程度以及其他行为因素为变量,运用统计学方法,以生存分析为视角,预测客户流失概率与可能性,能为企业提高客户关系管理提供有意义的决策依据。

关键词:客户流失;生存分析;客户价值

分类号:F426.722 **文献标识码**:A **文章编号**:1673-1395 (2018)06-0079-05

客户关系管理尽管不是一个全新的课题,但是由于客户在企业发展中的重要地位,因此对其研究一直没有中断过。姚博(2017)研究表明,公司高端 20%的顾客创造出公司 80%以上的利润,在企业资源有限的情况下,有效地预测客户流失,保持优质客户成为客户关系管理的重心。

目前客户流失管理方面的研究主要体现以下几个方面。

第一,客户流失动机的识别。主要研究是客户在什么情况下,客户会流失。通过分析客户的行为动机及购买频率等因素来预判客户流失的动向。

第二,分析客户流失的动因。主要研究是什么原因促使客户流失,是企业自身原因,还是客户消费者偏好发生改变,等等。

第三,防止客户流失的应对措施。主要探讨在企业资源有限的情况下,分析客户流失的动因,有针对性地进行挽留客户,进而提高客户的留存率。

通过对前人的研究整理发现,更多的研究体现在定性而非定量研究,有些虽是定量研究但在实务操作中不具有可操作性,更多的是停留在理论阶段层面。因为企业在评估客户对企业的价值时,不仅要考虑成本与效益的问题,还要考虑可操作性与外部环境等因素。因此,如何构建一个可操作的客户流失预测模型,对企业价值的提升具有十分重要意思。

一、客户流失的相关理论

(一)客户流失的界定

客户流失就是原客户不再购买原企业的产品或服务。Yeh I C(2009)将客户流失定义为“转换意愿”,客户流失就是指客户不再重复购买或终止原先使用的产品或服务。但是定义较为笼统,不够准确。例如客户不再重复购买没有时间限定,是一个月还是一年,另外客户不重复购买的行为与产品的特性也有关系。大件商品重复购买的周期比较长,而日常用品其重复购买周期就比较短。因此有些学者从定量的角度探讨客户流失的定义。张珠香(2018)认为当一个客户连续 3 个月没有在该企业进行任何消费,就是客户流失。林芳(2016)将流失的客户界定为已经流失指彻底停止消费企业所提供的产品或服务,同时作者还对客户流失进行简单分类:包括已经流失的客户和即将流失的客户[1~4]。即将流失的客

收稿日期:2018-10-18

基金项目:福建省教育厅社会科学研究项目(JAS170792)

第一作者简介:叶敏(1980-),女,福建福州人,讲师,硕士,主要从事经济管理研究。

户就是较之前对企业所提供的产品或服务消费变少,而已经流失的客户是已经开始向企业的竞争对手寻求替代品,但目前还没有完全断绝与公司的交易。也有学者王莹(2015)将客户流失按流失意愿分为自发流失、强制流失和预期流失。笔者研究的主要是自发流失。因为强制流失的客户实质上未能给企业带来效益,这样的客户对企业没有价值;而预期流失表明客户从根本上不再需要企业产品,是正常的退出机制,不会减少企业价值。

(二)客户流失预测方法

客户流失预测方法随着科学技术的不断提升,大数据应用的普及,其预测也越来越精确。目前对客户流失预测采用的主要方法是利用大数据,运用决策树、神经网络、遗传算法、生存分析、回归等,同时结合数据分析等软件操作,通过数据库的统计,分析客户购买时间,购买频率,购买数量,等等,预测客户流失动向。

常见的客户流失模型有二元结构模型和预测客户剩余生存期。二元结构模型就是将客户的流失分为两个维度:一个是客户流失的维度,一个是客户保留的维度。然后运用逻辑回归等相关方法建立模型,预测客户流失规律、时间及分布。客户剩余生命期就是利用企业客户数据库,分析客户在企业中保留期限,其目标就是建立模型评估客户流失的一种方法。

比较有代表性的研究成果有 Kisioglu P(2011)运用贝叶斯方法对电信公司客户流失行为进行预测。研究结论表明客户平均通话时间等因素是判断客户流失倾向的较为重要因素,为企业尤其是电信企业流失预测模型的改进提供有效指导。郑为益(2011)运用生存分析技术风险模型建立客户流失预测模型,分析客户流失的主要因素,为通讯运营商进行后续有针对性的客户营销方案提供重要理论决策依据。

(三)客户流失动因分析

企业的客户千差万别,企业处理客户关系也不尽相同。客户流失动因较为复杂,包括主观原因、客观原因、内部原因、外部原因,等等。目前研究没有得出一个统一的结论。余路(2016)从客户满意和客户价值的角度,分析客户流失的主要原因。研究表明当客户的价值受损或满意度降低,客户就会流失。也有学者从企业的角度研究客户流失,认为客户流失主要是企业未能有效重视客户管理,造成客户的购买意愿下降进而流失,并且这种单一客户的流失会导致其他客户购买意愿降低[5~7]。较为常见的情形就是企业未能妥善处理客户投诉,导致客户流失。

(四)客户挽留的理论研究

客户挽留实质上是企业的一种补救措施。如果企业在前期客户管理中能够有效满足客户需求,增加客户满意度,提高产品的附加值,不但不会造成客户流失反而会吸引更多的客户。因此客户挽留一定要做到有针对性的挽留,不能脱离企业与客户而无目的的挽留。较为有效的客户挽留要分析客户流失动因,分析动因产生的机理,同时结合企业成本效益前提下,评估挽留客户给企业带来价值与成本关系情形下,进而决定采用何种挽留措施。

目前关于客户挽留的研究主要是从定性和定量两个角度。定性研究通过研究客户流失动因出发,提出客户挽留建议。其主要优点在于简单可以行,不需要大量数据支撑,但是其缺点就是无法量化,不能准确的衡量企业挽留客户的成本效益比。相反定量研究就是利用大数据,采用一些数理统计模型进行分析,能够较好的测算出挽留客户的成本效益。但是其不足也非常明显就是需要大量数据支持,计算模型复杂,参数较多,结果受参数质量影响较大。孙树垒(2011)通过定量研究的方法建立客户挽留模型,提出客户保持对客户挽留的重要意义。[8]

二、生存分析理论

(一)生存分析的界定

生存时间原本是一个医学名词,是指某种疾病患者从开始患病到死亡所经历的时间跨度。而本文将生存时间界定为客户与企业初次购买到终止购买关系时间过程。

生存率就是客户留下的可能性,指客户经历 若干个时间个单位时段后仍与企业保持消费关系的可能性。流失率与生存率的关系是,流失率=1-生存率。

因此生存分析就是用来研究客户保留状态的规律。如客户挽留的时间分布特点,某一时间段内客户的挽留比率。其优点在于解决传统统计模型对数据要求过高的缺陷。

(二)生存分析模型

生存分析模型中最为重要的就是对生存函数的估计,而生存函数常用的估计方法,有参数法、半参数法等方法。参数法就是先对某种事件与时间的关系作出特定假定,并通过研究时间与对象之间的特定联系建立客户的生存函数 $S(t)$ 和时间 t 的关系。

在使用参数法对生存函数进行模拟估计是通常采用指数分布模型。指数分布是一种常用的概率统计分布,用来描述独立随机事件发生的时间间隔,反应时间与事件之间的相互变量关系。将变量关系用其概率密度函数进行表述:

$$f(t)=\begin{cases}\lambda e^{-\lambda}t>0\\0t\leqslant 0\end{cases}$$

其生存函数为:

$$S(t)=\begin{cases}e^{-\lambda t}t\geqslant 0\\0t<0\end{cases}$$

式中,参数 $\lambda>0$ 是指数分布的一个参数,数理上对其有明确的界定。指数分布的函数区间是[0,∞)。含参数 λ 的指数分布其均值和方差分别是 $\frac{1}{\lambda}$ 和 $\frac{1}{\lambda^2}$。

但是,实际中,如果某些参数无法获知的情形下,通常使用半参数法,就是模型中部分变量予以量化,部分变量定性分析。非参数法就是对客户保持与挽留的时间分布不作任何假设,直接对样本数据进行统计分析。半参数模型较参数模型的优点在于该模型不考虑研究对象的时间与事件之间的变量关系,允许观测对象在数据有缺失或截尾现象,并且可以全面观测变量与时间的长效关系。因此半参数模型对客户流失的预测,从客户生存角度无疑是一个比较好的方法。较为经典的半参数模型就是 Kotler(1999)提出的 Cox 比例风险模型。

该模型的数理表达式为:

$$h_i(t)=h_0(t)\exp(\sum_{j=1}^{p}\beta_j X_{ij})$$

式中,X_j 是随着时间推移,会对客户的生存结果造成影响的变量因素。$h_0(t)$ 是指自变量因素 $X_j(j=1,2,\cdots,P)$ 都处于某种特定状态下的特定系数,$h_0(t)$ 数字的确定与该函数特征状态有关。$\beta_j(j=1,2,\cdots,P)$ 成为 Cox 回归系数,是模型中的待定参数。

任何两个个体风险函数之比,即相对危险度可写为:

$$\begin{aligned}RR&=\frac{h_i(t,x)}{h_j(t,x)}\\&=\frac{h_0(t)exp(\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_px_{ip})}{h_0(t)exp(\beta_1x_{j1}+\beta_2x_{j2}+\ldots+\beta_px_{jp})}\\&=\exp[(\beta_1(x_{i1}-x_{j1})+\beta_2(x_{i2}-x_{j2})+\ldots\\&+\beta_p(x_{ip}-x_{jp})]\\&i\neq j,i,j=1,2,\ldots n\end{aligned}$$

RR 是相对危险度的一个结果,该比值在一定特定的状态下是一个特定数字,与时间无关,成为比例风险假定,简称 PH 假定。因此 β_j 的参考意义是:当自变量 X_j 每改变一个观测单位时,所引起的相对风险度的自然对数值。

三、客户流失预测模型的构建

笔者在前人的研究基础上构建客户流失预测模型,总体思路是从企业端提取客户交易数据,采用一定的数理统计方法,据不同群体的特征对客户进行分类,然后基于生存分析的视角建立客户流失预测模式。

客户分类中主要采用聚类分析方法。聚类分析是对分类对象按是否具有同一属性或类似属性的客户分为一类,在数理上主要通过客户的数量关系来表述,即不同分类主体之间的距离来度量分类对象是否有差异,这样有利于后续模型的构建及数据处理。在聚类分析中,笔者对不同客户群体运用下列数量指标进行分类:设 x_{ik} 为第 i 个对象的第 k 个指标,每个对象测量了 p 个变量,则对象 x_i 和 x_j 之间的距离(D_{ij})的定义为:

$$D_{ij}(q)=(\sum_{k=1}^{p}|x_{ik}-x_{jk}|^q)^{1/q}$$

式中 q 为大于 0 的正数。

同时也是基于生存分析视角探讨客户流失预测模型,因此对客户流失率也给予限定。客户流失率是指在一段时间内,客户流失数量与企业全部客户的比值。而生存率是指客户在一段时候后仍与企业保持购买关系或意愿的可能性,常用 p(x>t)表示。客户流失率、客户保持率的关系为:客户流失率+客户保持率=1,表明客户保持率越大,客户流失率也越小。

基于上述分析构建了基础生存分析视角的客户流失预测模型,模型如下:设观测从 0 时刻开始,记录样本总体的 n 个观测对象的生存数据 $t_1,t_2,\ldots t_n$ 为寿终数据时,记 $\delta_i=1$;若 t_i 为右删失数据时,记 $\delta_i=0$。$t_{(1)}\leqslant t_{(2)}\leqslant\ldots t_{(n)}$ 是 $t_1,t_2,\ldots t_n$ 的次序统计量,对应得到一系列 $\delta_{(i)}(1\leqslant i\leqslant n)$,用下列函数表示:

$$\hat{S}(t)=\begin{cases}1,t\in[0,t_a)\\\prod_{i=1}^{j}(\frac{n-j}{n-i-1})^{\delta(i)},\\\quad t\in[t_{(j)},t_{(j+1),}),\\\quad j=1,2,\cdots\cdots,n-1\\0,t\in[t_{(n),}\infty)\end{cases}$$

即设 n 是包含所有删失数据和非删失数据的观测对象总数,将 n 个对象生存时间观察值从小到大排序,使得 $X_{(1)} \leqslant X_{(2)} \ldots \leqslant X_{(n)}$,则有生存函数:

$$S(t) = \prod_{X(i) \leqslant t} \frac{n-i}{n-i+1}$$

式中 i 取遍所有满足 $X_{(i)} \leqslant t$ 的正整数,这里 $X_{(i)}$ 是非删失观察。

四、客户流失预测模型的实证检验

研究的数据来源于对福建省某大型商场 2014～2017 年的交易数据,根据其内部数据源,进行有效的数据筛选。为了保护调研对象的商业信息,对部分数据进行适当修正。样本总数定位在 2131 名客户,该样本均为有效样本,已考虑相关不合理数据。统计变量中购买金额用 M 表示,以元为单位;客户两次购买时间间隔用 T 表示,以天数为单位;客户两次购买时间间隔越长表示客户越容易流失,因此,笔者将客户流失界定为两次购买时间超过半年以上的;购买频率用 N 表示,单位用次数表示。

在对样本分析过程使用 SPSS 进行相关数据分析,样本的描述性统计见表 1、表 2、表 3。

表 1 样本描述统计

指标	N	极小值	极大值	均值	标准差
T	2131	3	1022	384.48	336.28
N	2131	1	132	12.72	22.25
M	2131	.120	22873.20	12478.80	33881.24
有效的 N(列表状态)	2131				

表 2 方差分析表

	聚类		误差		F	Sig.
	均方	df	均方	df		
购买金额	5564785.319	9	106544.287	2131	35.752	.000
购买频率	30027.456	9	248.662	2131	127.124	.000
间隔时间	3.786E22	9	52884696.124	2131	6785.230	.000

表 3 特征客户群

类别	TNM	人数	人数占比/%	购买总金额	金额占比/%
A	T↓N↑M↑	410	19.24	21785474	79.69
B	T↑N↓M↓	1721	80.76	5554426	20.31

从表 2 中我们可以看出 p 值(sig.)很小,因此可以初步购买金额、购买时间间隔和购买频率三个变量是影响客户分类的主要因素。上述通过 TNM 分类,将调研的客户群体分为如上两大类。

A 类客户群体:该类客户近期有购买,其购买的次数与其消费金额较大,该类客户购买时间间隔低于人均平均间隔时间 132 天,购买次数多于人均购买次数 12 次,购买金额大于人均购买金额 12478.80 元,人数 410 人,占比 19.24%,购买总金额 21785474 元,占比 79.69%。因此该类客户为企业的有效客户,可以持续为企业带来价值,企业应当重视对该类客户的维持与提升。

B 类客户群体:该类客户表现为近期几乎没有购买行为,或者购买次数较低,远远低于平均数。从表 3 中可以看出 B 类客户购买时间间隔高于平均客户购买的时间间隔 132 天,购买次数少于人均 12 次的购买次数,购买所消费的金额低于人均购买金额 12478.80 元,人数 1721 人,占比 80.76%,购买金额 5554426 元,占比 20.31%。因此该类客户属于企业流失的客户,企业可以通过后续的客户管理进行持续观察。

尽管笔者尝试建立一个定量预测客户流失模式,但是由于模式中涉及参数角度,尤其是与时间相关的变量,其输入值层级的高低,最终会影响模型的准确性,因此在参数的限定方面,研究做得还不够,模型的精准度需要提高。同时,在对客户分类与识别过程中,由于使用的数理方法不同,造成分类结构有一定的偏差,对后面模型结果会造成一定影响。因此,在后续的研究中尽可能以观察输入值为依据,提高参数的可控性,进而提高预测模型的精准度。

参考文献:

[1]姚博.客户流失预测模型研究及其应用[D].西北大学,2017.

[2]Yeh I C, Yang K J, Ting T M. Knowledge discovery on RFM model using Bernoulli sequence[J]. Expert Systems with Applications, 2009(3).

[3]张珠香.基于生存分析模型的电信客户流失研究[J].福州大学学报(哲学社会科学版),2018(1).

[4]林芳.基于决策树的客户流失模型的建立[J].赤峰学院学报(自然科学版),2016(21).

[5] Kisioglu P, Topcu Y I. Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey[J]. Expert Systems with Applications, 2011(6).

[6]郑为益.基于生存分析的客户流失模型研究[D].华南理工大学,2011.

[7]余路.电信客户流失的组合预测模型[J].华侨大学学报(自然科学版),2016(5).

[8]孙树垒.基于客户识别的客户保持决策模型与定价策略[J].管理学报,2011(10).

责任编辑 胡号寰 E-mail:huhaohuan2@126.com

(上接第 31 页)

不等于揭示了诗句的内涵,因此,必要的时候笺注者以“串讲”揭示大意。吴注“手提”二句、“淫僻”二句皆无法串讲(因释词有误)。熊注“淫僻”二句、“灭火”二句亦难以串讲(因典故有误)。赵注的优点在于几乎每句皆有串讲,但有的似乎引申太远,不能与原来的字词对应,如“淫僻”句“谓修身基于仁义,不事淫邪”。这个句子本身并没有什么问题。但“淫僻畏仁义”的“畏”字作何解释?李瑄将“淫僻”句理解为诗人“淫僻”所以“畏仁义”,将“行止”句理解为诗人的“行止”让“罔两”害羞。就过于突出了袁宏道的叛逆精神了。袁宏道有叛逆精神,思想也很驳杂,但一生仍在儒家界内,不可能在二十几岁时宣扬自己与仁义对立。分析失误的原因,主要是因为望文生义,如将淫僻解释为淫荡邪僻,将灭火解释为黑暗中,皆是望文生义所致。未能博览广搜以致错会典故亦属此列,如赵注对“无孔锤”的解释以及熊注对“灭火”的解释等。还有一种失误就是以注者的主观想法强加于诗句以致造成意义的扭曲,如对“淫僻”“行止”二句的解释。失误在所难免,关键是如何看待失误,如何一步一步地减少失误。只有以科学严谨的态度认真对待失误,才能使研究更上新的台阶。胡适有一个著名论断:“发明一个字的古义,与发现一颗恒星,都是一大功绩。”[12](P327~328)

当然,对袁宏道诗文的多种版本的注解在总体上是值得肯定的。以上罗列排比诸种问题并非对相关注解的完全否定。这些注解工作,筚路蓝缕,绝非易事,出现疏误,在所难免。上述所列注者,皆学界巨擘,他们的努力为袁宏道诗文研究普及以及“公安派”研究普及作出重要贡献。所列诸条,个别属于纰漏,有的尚须讨论。但无论如何,袁宏道诗文笺注这一基础工作还较为薄弱是个不争的事实。期待方家能作出全面、准确的笺注,从而将袁宏道及晚明“公安派”研究推上一个新台阶。

参考文献:

[1]钱伯城.袁宏道集笺校[M].上海:上海古籍出版社,1981.

[2]吴调公.公安三袁选集[M].武汉:湖北人民出版社,1988.

[3]熊礼汇.公安三袁[M].长沙:岳麓书社,2000.

[4]赵伯陶.袁宏道集[M].南京:凤凰出版社,2009.

[5]赵敏俐.中国诗歌研究(第 8 辑)[M].北京:中华书局,2011.

[6]何宗美.袁宏道诗文系年考订[M].上海:上海古籍出版社,2007.

[7]李健章.《袁中郎行状》笺证[M].武汉:武汉大学出版社,2012.

[8]钱伯城.珂雪斋集[M].上海:上海古籍出版社,1989.

[9]孟森.明清史讲义[M].北京:中华书局,1981.

[10]沈德符.万历野获编[M].北京:中华书局,1959.

[11]何宗美.袁宏道诗文系年刍议[J].文学遗产,2008(6).

[12]欧阳哲.胡适文集(2)[M].北京:北京大学出版社,1998.

责任编辑 周家洪 E-mail:zhoujiahong2004@163.com