基于数据挖掘的证券营业部客户流失分析

□王伟钧 [成都大学 成都 610106]

[摘 要] 针对营业部客户流失的影响因素数据给出了预测模型。根据新佣金政策下的证券营业部客户一年来的交易数据分析，提出了影响客户流失的各种特征因素的假设。进一步建立三个模型（逻辑回归、决策树和径向基神经网络），并从错分率和预测能力（提升值）两方面的模型性能评判标准来选出最佳模型，根据该模型找出最可能影响客户流失的各种预测因子，并进行深入分析，最后提出相应的对策。

[关键词] 流失，特征因素，模型，错分率，预测
[中图分类号] F2,TP391 [文献标识码]A [文章编号]1008-8105(2009)01-0018-05

引言

由于竞争对手行为对存在的客户关系的严重影响，客户生命周期正变得越来越短暂，这已给客户提供了极其多的选择机会。有不少客户成为企业的新客户的同时，又有大批客户流失。一些客户限制他们的选择变成面向的关系[1]，并有潜力变成长期客户[2]。另一些客户为了减少可能的转移成本，则成为分散在多个竞争企业中的客户[3]。在成熟的市场和竞争的压力时期，越来越多的企业真正认识到他们最珍贵的资产是存在的客户资源等[4]。

现在，证券企业和客户的关系并无协议限定，因而这些企业易遭遇客户的无声无息的（全部）流失，几乎没有给企业可以挽留他们的时间和机会，这将给企业造成不同程度的损失。另一方面，从长期来看，客户部分流失可能导致全部流失[2]，客户培养和保留比客户吸取更为重要[5]，保留客户能产生更多的利润。通过对美国的9个行业的调查研究显示，当客户保持率提高25%～85%时[6]，保留的客户能产生比新客户更高的收入和边际收益。客户保留、防止客户流失是确保企业的长期利润和成功的一项有价值战略[2]。

通过数据挖掘，组织能找到有价值的客户，预测未来的行为，使企业能提前采取行动，进行知识驱动的决策[7]。近十几年来发展起来的数据挖掘技术为解决上述问题提供了一种可行性解决方案[8]。本文将分析证券行业，以流失作为因变量，人口统计变量、行为变量、宏观环境变量和感知变量作为解释变量，通过逻辑回归、决策树和神经网络分别建立流失模型，选择最好模型来回答上述问题，并作相应的解释。最后提出相应的策略。

一、方法

（一）客户流失的内涵

对客户流失的定义不尽相同。大多数流失的定义为客户中断与公司的关系。Keaveney则认为流失转换意愿。文献[7]指出账户余额流失75%即为流失。文献[5]则认为忠诚和流失是一个问题的两个方面，指出：忠诚率＝1－流失率。由于客户流失的定义不明确，本文定义客户流失为客户的账户余额减少75%（部分流失）或中断与公司的关系（全流失，如销户等）。

（二）三个分类技术

数据挖掘定义为用统计算法发现数据中的模式和相关性的复杂数据寻找能力。数据挖掘技术用于收集对象（观察值），该对象是建立在变量上。为简化每种技术的描述，这里假设：有 N 个样品（观察记录数），每个样品有 k 个特征变量，即 $x_{i1},x_{i2},\cdots,x_{ik}$。$\overline{x_i}=(x_{i1},x_{i2},\cdots,x_{ik})$ 代表 k 个特征变量的一个矢量。y_i（$i=1,2,\cdots,n$）是二值相应变

[收稿日期] 2008－03－04
[作者简介] 王伟钧（1963－ ）男，成都大学计算机科学系副教授，在读博士.

量的观察值，只可以取值0或1。取值为1代表感兴趣的事件发生了，称之为“成功”。

1．逻辑回归（logistic Regression）

一个定性相应问题通常被分解成两值相应问题。大多数定性变量相应模型的基本元素是Logistic模型。Logistic回归模型根据一个拟合值来定义，被解释成事件在不同子空间中发生的概率：

$$\pi_i = P(y_i = 1) \quad （i = 1,2,\cdots,n）$$

拟合概率的logit函数表示时间发生的概率（成功）和事件没有发生的概率（失败）的比之的自然对数，表示为：

$$\text{logit}(\pi_i) = a + b_1 x_{i1} + b_2 x_{i1} + \cdots + b_k x_{ik} \quad （1）$$

式中，$x_i = (x_{i1}, x_{i1}, \cdots, x_{ik})$（$i$ 为观察记录序号，k 为观察变量的属性数）。一旦根据数据计算出 π_i。就得到每一个二值变量观察值的拟合值 $\hat{y}_i$。引入阀值 π_i，大于 π_i 则 $\hat{y}_i = 1$，小于 π_i 则 $\hat{y}_i = 0$。该模型具有非常显著的特点：1）可以求出事件发生的后验概率；2）很多类的分布都满足logit的基本假设[10]；3）易于使用，能提供快速的和健壮的结果；4）能非常好地用于多维及OLAP数据；5）能一致地处理数值和非数值数据。

2．径向基函数神经网络（RBF）

人工神经网络经常为取得比（统计）分类技术具有较高预测性能而创建的[9]。普通径向基函数网络是有单一隐藏层的前向神经网络。选择它主要是因为在输入变量空间中有邻接结构，而且需要说明它。隐层单元的转移函数采用径向基函数，以对输入层的激励产生局部响应。RBF克服了传统前向神经网路的许多缺点，它训练快，在训练时不会发生振荡，也不会陷入局部极小。RBF具有很好的通用性，只要有足够的隐层神经元，RBF就能以任何精度近视任何连续函数。最常用的径向函数为高斯函数：

$$g_0^{-1}(E(y)) = w_0 + w_1 \exp(-w_{01}^2 (x - w_{11})^2) \quad （2）$$

式中，$g_0^{-1}(.)$为输出激活逆函数，w_0 为偏度，w_1，$w_2 \cdots w_n$ 分别为对应的输入变量的权重，w_{0i}^2 为反比于基函数宽度，$(w_{1i}, w_{2i}, \cdots, w_{ni})$ 为权重。其可调节参数有两个，即中心位置和方差，网络的可调参数有三组，即各基函数的中心位置。方差和输出单元的权值，通常选择隐层的节点数为训练样本个数，每个节点都有一个径向基函数的中心向量，该中心向量为训练样本的输入向量。神经网络的参数，通过将验证数据集合上错误分类率降低到最小程度而得到。

3．决策树（Decision Tree）

决策树学习是以样本为基础的归纳学习方法。决策树为解决分类任务而变得非常普遍，因为它能处理预测因子，而这些因子以不同的度量水平（包括名字变量）来度量，同时还因为决策树易于使用和解释[10]。然而，决策树也有诸如缺少健壮性和次优的性能。基于决策树的学习算法在学习过程中不需要用户了解很多背景知识，只要训练样本能够用属性-值的方式表达，就可以使用该算法来学习。

4．三个模型性能比较

我们评价这些模型的性能，用错分率（Misclassification Rate）和提升图。这两个度量方法在很多文献中普遍使用[11]。错分率在方法研究中特别用于评价几个竞争的性能。提升图把验证数据集中的观测数据根据其分数以升（或降）序排列，分数是基于训练集估计的相应事件（成功）的概率。把这些分数再细分成10分位点，接着对验证集中的每个10分位点计算和图示成功的观测概率。如果这些成功的观察概率和估计概率具有相同的顺序(升序或降序)，那么模型是有效的。一个模型的提升图常与一个基本线比较，此时概率估计不是通过模型得到的，而是通过对观测的成功概率求均值得到。

二、研究的数据

为了实证分析，我们使用了位于西南某城市中心地带的证券营业部的证券交易数据。该营业部隶属于某一具有经纪（A股和B股等）、自营和投资银行等业务的大型券商。自2006年10月1日，沪深两地实行浮动佣金制，只规定上限，各券商可以根据自身服务质量和服务对象实施具体的佣金政策。为了反映新的佣金政策下的客户的流失性分析，我们取得有22863个客户包括2006年10月至2007年12月的数据。把2006年10月1日作为基准日，2006年10月~2006年12月数据作为政策前的客户基准数据，2006年12月～2007年12月作为客户流失研究的数据。面向客户流失的主题创建了数据仓库（集市）。按照下列要求从数据仓库中删除了部分客户的数据：

1）在2006年10月1日前已销户的；

2）在2006年10月～2007年12月期间，一直没交易并且票现合计（资金余额与证券市值之和）保持在1000元之内的。

通过上述处理，数据集中最后包括8028个客户

的数据，其中，男性占总数的54.19%，女性占45.81%，流失客户的总数为438人。在以下模型中我们将随机地取67%作为训练集数据，33%作为验证集数据。

三、预测因子与假设

（一）概述

定义1 如果基准日后1年的平均票现合计余额比基准日前的平均值流失75%，则流失为1，否则为0。

另外，现有文献对客户的流失预测主要是围绕四类变量来进行的。所以假设客户的流失至少是由人口统计变量、行为变量、感知变量和宏观环境变量之一决定的。从数据库中可以得到人口统计变量、行为变量和宏观环境变量。

（二）客户人口统计预测因子与假设

1）Age 文献关于年龄对流失的影响结论不一致。这里假设年龄对客户流失可能有影响。

2）Gender 相关文献关于性别对流失的影响结论不一致。通过探索性数据分析，这里对流失的影响不大，所以只假设性别对客户流失影响不显著。

3）Distance 未找到客户离商家距离远近对流失的影响的文献。但在客户细分研究中[12]，指出该变量与购买行为有关。通过探索性数据分析，似乎该变量与流失有关。然而，该地区营业部数量多且集中，客户更喜欢在离住家近的券商处交易。可能该变量与客户的流失性有关，所以这里假设距离对客户流失有影响。

（三）客户行为预测因子

1．RMF类 在这里该类变量包括：Recency（最后一次交易距观察日的天数）。假设Recency越小，客户的流失性越小；反之客户的流失性越大。

2．B_Stock 由于开通B股业务的券商不多，因而客户更愿意在有该业务的券商处交易，流动会增加成本。这里假设开通B股业务的客户，将更愿意继续在原券商处交易，即不易流失。

3．Rmop 资金类别一方面反映资金来源渠道，另一方面反映资金存取模式。由于有关国有企业买卖股票的要求，该变量可能决定资金使用期限，从而影响客户的流失性。所以假设资金类型可能影响客户的流失性。

4．LOR 客户关系长度可能反映客户与券商的牢固程度，所以假设客户关系长度越长，则客户的流失性越小。

5．ReturnLoss 按照直观认识，认为客户买卖证券盈亏程度，将可能不同程度地影响客户的流失性。因而，假设亏损的客户比盈利的客户更易流失，盈利少的客户比盈利多的客户更易流失。

6．InAssignStock、InTranStock和InFund 这3个变量反映客户资金或股票从其他渠道流入，可能说明客户对该券商有一定的偏好，因而，假设该变量越大，客户流失性越小，反之越大。

7．OutAssignStock,OutTranStock和OutFund分析假设撤销指定交易,转（出）托管变量或取（转）出资金越大，客户流失性越大；反之越小。

8．AVG_OCCUR_BALANCE 平均交易额。假设平均交易额越大，流失的可能性越大。

9．BALANCE_ATTR 支准点前3个月和后12个月的平均资金余额差。假设该变量值与流失性有关。

（四）宏观环境预测因子

1．MarketReturnRate 市场收益率变量反映客户关系期间（或观测期内），证券市场的繁荣与否。因而假设MarketReturnRate值越高，客户将越不易流失，反之则易流失。

2．Promotion 这里促销方式有三种。有关文献指出促销与流失成反相关关系[13，14]。可以想象促销强度和方式不一样，可能对客户的流失有不同的影响。这里假设无促销的客户将更可能流失，而通过赠送交易设备促销的客户比返佣金的客户更易流失。

（五）预测因子的变换

为了改善3个模型的性能，对大多数预测因子作了变换，产生新的变换（略）。

四、结果与解释

（一）三个模型性能

通过SAS软件，三个模型运行的结果在表1中。

从表1的错分率看，逻辑回归无论是训练数据还是验证数据均较低，也就是说该模型在分类上要好于其他两个模型。现在再看看三个模型的预测能力比较（表2）。从表中看出在预测能力上，逻辑回归最好，神经网络其次，决策树最差。但逻辑回归与神经网络几乎相同，只是稍优一些。

表1 错分率的概括比较

模型	训练数据	验证数据
逻辑回归	0.0056	0.0087
决策树	0.01	0.01
神经网络	0.0056	0.0094

表2 在各百分位处模型的累计提升值（基线响应率为5.456）

模型	累计提升值									
	10	20	30	40	50	60	70	80	90	10
逻辑回归	9.72	4.93	3.29	2.47	1.97	1.65	1.42	1.24	1.11	1
决策树	8.97	4.56	3.08	2.34	1.89	1.59	1.38	1.22	1.10	1
神经网络	9.66	4.93	3.29	2.47	1.97	1.64	1.41	1.24	1.11	1

（二）逻辑回归模型选定的预测因子

逻辑回归模型根据优势率选定的预测因子与优势率情况如下：

表3 选定的逻辑回归模型预测因子的优势率结果

变量	优势率
Age	1.035
AVG__GRK	1.445
BALA_7V3 01:low--306.2 vs 03:0-high	0.429
BALA_7V3 02:-306.2-0 vs 03:0-high	2.156
CLOS_YI6 01:low-0 vs 02:0-high	0.000
CURR_GD4 01:low-0 vs 03:1000-high	0.209
CURR_GD4 02:0-1000 vs 03:1000-high	2.036
PRE3_1IU	1.579
PRE3_QK2	4.685
Promotion 0 vs 2	9.049
Promotion 1 vs 2	43.419
PST1_5O1	0.200
PST1_7MX	0.315
YQ1_OQ85 01:low-0 vs 02:0-high	4.720
STOC_9P0 01:low-783.6 vs 03:205.7-high	0.000
STOC_9P0 02:-783.6-205.67vs03:205.7-high	0.830
INFU_OOE 01:low-0 vs 02:0-high	6.657
Satisfaction -1 vs 1	4.515
Satisfaction 0 vs 1	0.003

五、结论与策略

（一）结论

实证结果显示出三个模型（逻辑回归、决策树和神经网络）对客户流失性的错分率和预测能力上都有好的结果，但逻辑回归模型最好。通过逻辑回归模型在 SAS 上运行的结果表明：平均交易量、支准点前 3 个月和后 12 个月的平均资金余额差、票现合计余额、基准日前的平均资金额、基准日前的证券市值、支准点后 12 个月资金额、支准点后 12 个月股票市值对流失性具有特别显著的作用；促销方式为送交易设备和买卖盈亏也非常显著地影响客户的流失性；年龄对流失性的作用较显著。更易于流失的客户应该满足：

平均交易量大的、支准点前 3 个月和后 12 个月的平均资金余额差在（–306.2,0）、票现合计在（0，1000）的客户，基准日前的平均资金额大、基准日前的证券市值大、支准点后 12 个月资金额小、支准点后 12 个月股票市值小、买卖亏损大、没有存入过资金、年龄较大和对投资收益率不满意。

（二）策略

1．给客户以亲切感，进行情感投资。与客户保持互动式沟通，使企业与客户建立一种牢固的联系，对这种联系的维持进行情感投资，使顾客和企业休戚相关。

2．给客户更多方便和更多选择，企业在满足顾客需求的基础上，要建立一系列便利条件，使顾客可随时随地获得服务，使客户获得更高的满意度。

3．多举办高水平的证券投资咨询报告，一方面，树立客户的正确投资理念，提高客户抗风险的能力；另一方面给客户提出较有指导性的参考性报告。

4．提供个性服务，更有效地满足顾客需求。企业应把每一个有价值的顾客当作永恒的服务对象，而不是一次交易对象。当今时代是个性化需求的时代，企业必须抢占网络先机，在充分了解顾客需求的基础上，充分分析客户的持仓情况，提出具有针对性的对策。

5．提供快速、有效服务，面对瞬息万变的市场作出立即反应，要做到满足顾客需求，建立关联关系，企业必须建立快速反应机制，提高反应速度和回应力。

6．取消交易通讯费，降低客户交易成本，减少流失，同时可能吸收原流失的客户；

7．密切注意亏损客户，分析亏损原因。提供专业的证券投资理念指导，避免过度频繁交易；

8．完善促销手段。从竞争对手那里学习促销手段，避免让客户出现“优惠条件差”的感觉。同时，创新或改进促销手段，如多举行故事沙龙、业余活动等。

进一步的工作还需对客户进行满意性的调查研究，以便找出客户流失的真正原因。此外，还应对客户进行细分，以采取更具针对性的策略。

参考文献

[1] SHETH J N, PARVATIYAR A. Relationship in consumer markets: Antecedents and consequences[J]. Journal of the Academy of Marketing Science 1995, 23 (4): 255-271.

[2] BUCKINX W, DIRK Van den Poel .Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting[J]. European Journal of Operational Research 2003, 164: 252-268.

[3] DWYER R F.Customer lifetime valuation to support marketing decision making[J]. Journal of Direct Marketing 1997, 11 (4): 6-13.

[4] ATHANASSOPOULOS A D. Customer satisfaction cues to support market segmentation and explain switching behaviour[J]. Journal of Business Research 2000, 47 (3): 191-207.

[5] HWANG H, JUNG T, SUH E. An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry[J]. Expert Systems with Applications, 2004,26(2):181-188

[6] REICHHELD F F. Learning from customer defections [J]. Harvard Business Review 1996, 74 (2): 56-69.

[7] 王伟钧，杨承师，杨晋浩．企业的知识发现与经营模式重组[J]．决策咨询通讯，2002,(50):25-26。

[8] RUD O P. 数据挖掘实践[M]. 北京：机械工业出版社，2003。

[9] CHRIS RYGIELSKI A, WANG B J C, DAVID C, etal. mining techniques for customer relationship management[J]. Technology in Society 2002, 24: 483-502

[10] BAESENS B, VIAENE S, VAN den POEL D, et al. Bayesian neural network learning for repeat purchase modelling in direct marketing[J] European Journal of Operational Research 2002, 138 (1),:191–211.

[11] ANDERSON J A. Logistic discrimination. In: Krishnaiah, P.R., Kanal, L.N. (Eds.)[J], Handbook of Statistics, 1982,(2): 169- 191.

[12] RUIZ J P, CHEBAT J C. Another trip to the mall: a segmentation study of customers based on their activities[J]. Journal of Retailing and Consumer Services 2004, (11): 333-350.

[13] DUDOIT S, FRIDLYAND J, SPEED T P. Comparison of discrimination methods for the classification of tumors using gene expression data[J]. Journal of the American Statistical Association 2002, 97 (457): 77-87.

[14]余世雄．导致客户流失的五个原因[J]．江苏企业管理.2007,(9):30-31.

Analysis of Customers Churn and Data Mining from A Securities Business Department

WANG Wei-jun
(Chengdu University Chengdu 610106 China)

Abstract The paper builds a forecast model with factors to affect customers' churn. It presents some hyphothesises by these factors from the data of a securities business department. and builds three models (logistics, decision and nural network) and selects the best one from the above models by both disclassification and forecast ability. Moreover, it finds some factors which affect customers churn with maximum probability and presents some policies to prevent customers from churn.

Key words churn; model; disclassification; forecast

编辑 范华丽