DOI: 10.19327/j.cnki.zuaxb.1007-9734.2016.05.014

基于 logistic 模型的证券公司客户流失预警分析

郑宇晨，吕王勇
（四川师范大学 数学与软件科学学院，四川 成都 610068）

摘 要： 伴随着中国经济的高速发展和经济全球化的不断加深，客户流失问题比争夺客户更需要证券公司的高度关注。文章从反映客户交易情况的指标出发，采用 K－均值聚类获取客户流失状态；接着通过6种逐步回归方法进行变量筛选，并建立 logistic 客户流失预警模型；再对模型的泛化能力进行检验并基于证券公司的业务特点给出分析。研究结果表明：反映客户交易活跃度的指标是证券公司实施客户流失预警的关键，进而为证券公司有针对性地挽留客户提供有效的方法和可行的建议。

关键词： 客户流失预警模型；logistic 回归；数据挖掘；证券公司
中图分类号： F830.9 **文献标识码：** A **文章编号：** 1007－9734(2016)05－0080－09

一、引 言

改革开放以来，中国证券行业空前发展，来自国内外券商同行以及银行的多重竞争压力也接踵而至。有研究表明，相比提高市场占有率、扩大经营规模，减少客户流失对企业来说更具吸引力。券商每减少5%的客户流失，就能使盈利水平提高25%到85%；大多数新客户给公司创造的利润低于稳定的老客户。因此，保留住客户，对客户流失前的征兆及时预警，对于提高公司的竞争力有举足轻重的战略意义。

客户流失可以被定义为因为企业各种营销手段的实施导致客户与企业终止业务关系的现象。客户流失分析，旨在用数据挖掘为代表的方法，分析反映客户历史交易行为的数据，提取有流失风险的客户行为特征，将其应用于客户关系管理，改进或调整营销手段来实现挽留客户的目的。

客户流失预警研究领域如今成果颇丰。仲继(2014)[1]针对电信运营商的老客户保留问题，通过对客户流失原因的分析，将客户区分并给出不同的流失标准，分别用 C5.0 决策树、支持向量机、C&T 决策树、logistic 回归和神经网络分别建模预测，并最终通过增加一个置信区间的方法提出融合模型，降低了预测风险。姜晓娟、郭一娜(2014)[2]研究相同的问题，考虑客户流失数据正负样本不对称性且规模庞大的特点，对各个数据库增加权值参数；通过加权聚类，取得了较好的预测效果。王建仁(2015)[3]针对电信行业客户流失问题，提出将信息融合和多种数据挖掘方法相结合的融合模型，使模型的预测精度有了质的提高。杨孝成(2014)[4]对移动通信用户采用聚类算法，并以此为依据建立流失预警的决策树模型并设计了用户流失预警的基本结构，已经形成了较完备的算法和一定的实用价值。

然而，客户流失预警领域在证券行业的研究几乎是空白；前人的研究几乎都停留在宏观的证券公司客户关系管理上，且以方法论式的建议居多。仅王卉(2008)[5]针对证券公司的客户流失问题，建立服务质量的六缺口模型，给出了服务失败后的具体补救措施，并用实证数据对补救措施给出评价。杜修平(2009)[6]提出了影响证券公司客户流失的特征因素，用决策树方法构建了证券行业客户流失分析的 RFM－ROI 模型，并给

收稿日期：2016－06－17
基金项目：教育部人文社科规划项目(12YJA630197)
作者简介：郑宇晨，男，安徽蚌埠人，硕士，研究方向为应用统计、互联网金融。
吕王勇，女，副教授，博士，研究方向为应用统计。

出了剪枝的阀值，获得了 80.7% 的预测准确率和较强的实用性。吴斌(2013)[7] 针对证券公司的客户流失问题，结合证券经纪业务的特点，从资产累计流出量的角度选择变量建立 logistic 回归预测模型，获得了较理想的 K－S 值、提升度和一定的捕获率。

根据于彩娴、赵治荣(2013)[8] 对银行业的研究，针对不同的行业、数据库，应该在所处环境下探索最优的模型，没有一种模型的预测效果总是优于其他模型。故于彩娴、赵治荣等分别用 logistic 回归、决策树、人工神经网络、决策树和 logistic 回归的融合模型进行对比建模；结果表明，logistic 回归的提升性最高，决策树最低，决策树和 logistic 回归的融合模型次低。Verbek(2012)[9] 通过对 11 个数据集的研究比较，也说明模型的有效性因模型的检验方法和数据集而异。本文对证券公司客户交易数据进行筛选和 logistic 回归建模研究，并提出有行业针对性的建议，以期在相当程度上为证券公司的决策者提供参考。

二、logistic 模型

(一) 模型的建立

logistic 回归(logistic regression)模型属于非线性概率模型，是探究二分类观察结果与影响因素的定量关系最常用的模型之一。因变量是目标分类字段，偏回归系数解释为自变量的单位变化引起因变量变换后的平均变化。客户流失是典型的二分类问题：客户流失(1)，客户正常投资(0)。现实中，离散型模型对于捕捉数据的本质、解释和说明投资者行为更为有效，且离散型模型以最大化概率作为主要估计方法，保证了参数估计的一致性和有效性[10]。

对于证券公司的每一位客户，把其投资状态定义为，其中 $Y=1$ 代表客户流失，$Y=0$ 代表客户正常投资；根据客户的各项交易数据构造向量 $X(X_1,X_2,\cdots,X_n)$，建立 logistic 回归模型如下：

$$P \overset{def}{=} P(Y=1\mid X)=\frac{e^{\alpha+\beta_1X_1+\beta_2X_2\cdots\cdots+\beta_nX_n}}{1+e^{\alpha+\beta_1X_1+\beta_2X_2\cdots\cdots+\beta_nX_n}} \quad (1)$$

$$P(Y=1\mid X)=\frac{1}{1+e^{\alpha+\beta_1X_1+\beta_2X_2\cdots\cdots+\beta_nX_n}} \quad (2)$$

其中，$P(Y=1\mid X)$——客户流失的概率，$P(Y=0\mid X)$——客户正常投资的概率，X_i——自变量(解释变量)，本文是指客户在证券公司营业部的交易指标；β_i——各自变量的偏回归系数，代表对相应自变量的贡献，α——截距项。

(二) 模型的理论基础

为了与数学上习惯的表达相一致，将(1)式等价地改写为如下形式：

$$y=\frac{e^{\alpha+\beta_1X_1+\beta_2X_2\cdots\cdots+\beta_nX_n}}{1+e^{\alpha+\beta_1X_1+\beta_2X_2\cdots\cdots+\beta_nX_n}} \quad (3)$$

简单变形得：$\log it(y) \overset{def}{=} \ln\frac{y}{1-y}=\alpha+\beta_1X_1+\beta_2X_2\cdots\cdots+\beta_nX_n$ (4)

这里因变量与自变量 $X_1、X_2\cdots\cdots X_3$ 建立的回归方程就是 logistic 模型。这里把

$$y=\frac{1}{1+e^{-x}} \quad (5)$$

称为 logistic 函数。其中，$x \overset{def}{=} \alpha+\sum_{i=1}^{n}\beta_iX_i$

一般化的 logistic 函数形如：

$$y=\frac{K}{1+ae^{bx}} \quad (6)$$

K 代表承载能力或最大容量，反映系统内事物的饱和状态。参数 b 表示最大可能相对增长率，参数 a 的值由 K 与 y 的初始值 y_0 的比值来确定[11]。

对(6)式两边求导，得一般 logistic 函数的微分方程形式：

$$\frac{dy}{dt}=by\left(1-\frac{y}{K}\right) \quad (7)$$

(6)式是(7)式的特解，也是解析解。该曲线的基本解析性质如下：

(1) 单调性：严格递增；

(2) 渐近线 2 条，分别是 $y=0$ 和 $y=K$；

(3) 唯一拐点 $x^*=\frac{\ln a}{b}$ 满足 $y(x^*)=\frac{K}{2}$。

$\log it(y)$ 称为 logistic 变换，作用如下：

$$\lim_{x\to-\infty}y=\lim_{x\to-\infty}\frac{1}{1+e^{-x}}=0,$$

$$\lim_{x\to\infty}y=\lim_{x\to-\infty}\frac{1}{1+e^{-x}}=1 \quad (8)$$

通过求导可知 logistic 函数(5)在实数域 R 上单调递增，故 $y\in(0,1)$。

$$\lim_{y\to0}\log it(y)=\lim_{y\to0}\ln\frac{y}{1-y}=-\infty,$$

$$\lim_{y\to1}\log it(y)=\lim_{y\to1}\ln\frac{y}{1-y}=+\infty \quad (9)$$

这样，经过 logistic 变换，因变量的取值范围从(0,1)变换到$(-\infty,+\infty)$。

(三) 参数估计与假设检验

1. 参数估计。经典方法是最小二乘法。明显的，(4)式与多元线性回归形式相同，故可采用

相同的方法估计参数 α 和 β_i。根据估计后的方程，代入解释变量的观测值，即可得两类客户的流失概率。

2. 模型整体显著性检验。作用：检验自变量（指标）全体对因变量的影响是否有统计意义。常用的检验方法有似然比检验、Wald 检验和计分检验，其中似然比检验的结果最可靠，后两者可靠性相当。本文采用似然比检验。

似然比检验：通过分析模型中变量的变化对似然比的影响，来检验自变量的增加或减少是否对因变量产生统计意义上的显著影响。零假设和备择假设如下：

$$H_0: \beta_1 = \beta_2 = \cdots \beta_m = 0 \quad (10)$$

$$H_1: \text{各 } \beta_j (j = 1, 2 \cdots, m) \text{ 不全为 } 0 \quad (11)$$

检验统计量：

$$G = -2 [\ln(L_{k-1}) - \ln(L_k)] \quad (12)$$

其中，$\ln(L_{k-1})$ 是不包含检验变量时模型的对数似然值，$\ln(l_k)$ 是包含检验变量时模型的对数似然值。当 H_0 成立时，$G \sim \chi^2(n)$。

3. 回归系数的显著性检验。作用：检验单个自变量（指标）的偏回归系数与 0 是否有显著差异。主要是 Wald 检验，目前主流统计软件（SPSS、R、SAS 等）均采用此方法。零假设和备择假设如下：

$$H_0: \beta_j = 0 \quad (13)$$

$$H_0: \beta_j \neq 0 \quad (14)$$

检验统计量：

$$\chi^2 = \left(\frac{b_j}{S_{bj}}\right) \quad (15)$$

其中，b_j 是第 j 个自变量（指标）偏回归系数的估计值，S_{bj} 是 b_j 的标准误差。当 H_0 成立时，假定系数统计量服从正态分布，则 $\chi^2 \sim \chi^2(1)$。

4. 变量筛选的检验统计量。变量筛选的目的是将偏回归系数在统计意义上显著的自变量选入模型，不显著的自变量剔出模型。要特别说明的是，在 logistic 模型中，变量筛选不采用 F 统计量，而采用似然比统计量、Wald 统计量和计分统计量之中的一个。本文采用 Wald 统计量。

三、实证分析

（一）指标的初选及介绍

本章的实验数据基于某证券公司营业部部分客户的交易数据，时间跨度为 2013 年 1 月 1 日至 2013 年 12 月 31 日。抽取 4 155 个样本点，并将训练样本选为 3 500 个，测试样本选为 655 个。为了便于研究，本文将用户状态做二分划：流失和正常投资（说明：流失风险较高的客户即视为流失，其他视为正常投资）。

选择合适的解释变量是客户流失预测建模的重要环节。这里，原始变量选择鲜有研究的证券公司客户交易指标。因为移动通信、银行等领域的客户流失预警建模成果已经非常丰富，但缺乏在证券领域的相关研究可以借鉴，所以针对本文的数据特点选择原始变量见表 1：

表 1 原始变量介绍

编号	指标名称	指标含义	单位	取值范围	反 映	计算公式
1	周转率	客户买卖股票的频率	%	0 ~ 825%	客户交易活跃度	2013 年内股票的成交量/资产 x100%
2	平均持股时间	平均每只股票的持有天数	天	1 ~ 250	客户交易活跃度	2013 年内持有各只股票的时间求和/持股支数
3	日均仓位	平均每天客户实际投资资金占实有投资资金的比例	%	0 ~ 1	客户的交易习惯和风险承受能力	2013 年平均每天实际投资资金/实有投资资金
4	持股分散度	平均每天持有的股数	支	1 ~ 48	客户进行分散投资的程度	2013 年累计持股支数/持股天数
5	最长连续无交易时间	最长连续没有进行股票交易的天数	天	0 ~ 245	客户交易活跃度	——
6	普通账户交易量	客户的证券成交数量	支	0 ~ 17967	客户交易股票的盈利情况及对证券公司的贡献程度	——
7	最大上涨率	客户持仓个股的最大上涨率	%	0 ~ 805	客户的择股能力	max【某支上涨股票 2013 年 12 月 31 日股价 - 2013 年 1 月 1 日股价）/2013 年 1 月 1 日股价】
8	最大下跌率	客户持仓个股的最大下跌率	%	0 ~ 97	客户的择股能力	max【某支下跌股票 2013 年 12 月 31 日股价 - 2013 年 1 月 1 日股价）/2013 年 1 月 1 日股价】
9	投资收益	客户证券投资的收入减去损失的净收益	元	- 1951235 ~ 1876805	客户的整体投资水平	投资收入 - 投资损失

注：以上指标的时间跨度均为近一年（2013 年 1 月 1 日至 2013 年 12 月 31 日）.

（二）变量筛选

在变量筛选之前，首先进行数据预处理。第 172 个样本点的日均仓位数据缺失，用剩余有效样本的均值 0.8174 替换。因为 9 个原始变量观察值的量纲、数量级有很大差异，以下采用 Z－score 标准化（zero－mean normalization）方法对数据进行标准化处理，转换公式如：

$$x_i^* = \frac{x_i - \bar{x}}{\sigma} \quad (16)$$

1. 初步判断：简单相关系数检验。针对全部 4155 个样本点，原始变量的相关系数计算结果如表 2：

表 2　相关系数检验

	周转率	平均持股时间	日均仓位	持股分散度	最长连续无交易时间	普通账户交易量	最大上涨率	最大下跌率	投资收益
周转率	1.000								
平均持股时间	－0.419	1.000							
日均仓位	－0.050	0.237	1.000						
持股分散度	－0.101	0.206	0.111	1.000					
最长连续无交易时间	－0.370	0.550	0.006	－0.110	1.000				
普通账户交易量	0.368	－0.182	－0.062	0.090	－0.225	1.000			
最大上涨率	－0.103	0.113	－0.013	－0.019	0.101	－0.004	1.000		
最大下跌率	－0.019	0.145	0.185	－0.037	0.076	－0.045	0.412	1.000	
投资收益	－0.049	－0.022	－0.095	0.006	0.029	－0.068	0.106	－0.033	1.000

注：以上指标的时间跨度均为近一年（2013 年 1 月 1 日至 2013 年 12 月 31 日），为了使表达更加精炼，表 2 及对其的分析均省去“近一年”的表述.

由表 2 知，原始变量间相关系数高于 0.3（即线性相关显著）的结果有 5 个，分别是：平均持股时间和周转率、最长连续无交易时间和周转率、最长连续无交易时间和平均持股时间、普通账户交易量和周转率、最大下跌率和最大上涨率。最大的相关系数绝对值接近 0.55。而且，因为样本量较大，即使是因为随机因素的影响也会增大变量间的差异，从而导致相关系数较低。

所以，有必要进行变量筛选。本文采用 logistic 逐步回归的方法，因此，需要先获得因变量：客户流失状态。

2. 因变量的假设及获得。模型假设：因为客户流失状态数据是证券公司的商业机密，无法获得。因此，本文在实证部分假设各项指标表现“较激进”的客户为流失客户，交易活跃度低，投资能力差，流失风险高，即对证券公司的贡献小，Y 值取为 1；各项指标表现“较稳健”的客户为正常投资客户，交易活跃度高，投资能力好，流失风险低，即对证券公司的贡献大，Y 值取为 0。

本文采用 K－均值聚类（K－means cluster）判断客户所属的流失状态，对客户分类，并将分类数 K 定为 2。为了模型评价的需要，全部 4 155 个样本点均参与聚类。限于篇幅，这里不完整展示聚类结果（因变量的取值结果）。最终两种分类的聚类中心结果见表 3：

表 3　最终聚类中心

	聚类	
	1	2
近一年周转率	0.305236732	－0.477156
近一年平均持股时间	－0.61783926	0.965826455
近一年日均仓位	－0.2370511	0.370566
近一年持股分散度	－0.12421906	0.194183279
近一年最长连续无交易时间	－0.47088271	0.736099183
近一年普通账户交易量（万）	0.157584975	－0.24634197
近一年最大上涨率	－0.2058135	0.321734368
近一年最大下跌率	－0.25040773	0.391445524
近一年投资收益（元）	－0.00448579	0.007012338

由表 3 得，第 1 类的聚类中心为（0.305，－0.618，－0.237，－0.124，－0.471，0.158，－0.206，－0.250，－0.004），第 2 类的聚类中心为（－0.477，0.966，0.371，0.194，0.736，－0.246，0.322，0.391，0.007）。

由聚类结果可知，第 2 类客户具有以下特点：

近一年普通账户交易量（单位万）为负值，其他指标均为正值。并且，除了近一年投资收益（元）接近于 0 外，其他指标均显著不为 0。近一年周转率、平均持股时间、日均仓位指标说明第 2 类客户的交易活跃度较低，即参与度较低；近一

年普通账户交易量较小说明第2类客户的盈利状况较差、对证券公司的贡献也较小; 近一年最长连续无交易时间、持股分散度、最大上涨率、最大下跌率、投资收益(元)指标说明第2类客户资产较雄厚,愿意参与“高风险,高收益”的投资。综上,第2类客户的流失风险较高,被定义为本文中的流失客户,Y值取为1。与此相对应,第1类客户为正常投资客户,Y值取为0。

为了检验分类的合理性,用方差分析来检验两个类别之间是否有显著差异,结果如表4:

表4 ANOVA

	聚类		误差		F	Sig.
	均方	df	均方	df		
近一年周转率	605.157	1	0.855	4153	708.18	0.000
近一年平均持股时间	2479.394	1	0.403	4153	6148.866	0.000
近一年日均仓位	364.988	1	0.912	4153	400.05	0.000
近一年持股分散度	100.224	1	0.976	4153	102.677	0.000
近一年最长连续无交易时间	1440.191	1	0.653	4153	2203.955	0.000
近一年普通账户交易量(单位万)	161.296	1	0.961	4153	167.772	0.000
近一年最大上涨率	275.133	1	0.934	4153	294.577	0.000
近一年最大下跌率	407.277	1	0.902	4153	451.44	0.000
近一年投资收益(单位元)	0.131	1	1	4153	0.131	0.718

从表4的分析结果可见,除了第9个变量近一年投资收益(单位元)的P值高达0.718,说明两个类别在该指标上没有显著差异; 其他变量的P值均接近于0,说明前8个指标对分类结果的产生均有高度显著的影响。因此,把4 155个客户样本点按流失状态分成2类是合理的。

3. 解释变量的筛选: logistic逐步回归。因为SPSS21.0将客户分类标记为“1”、“2”,而被解释变量(因变量)Y的取值为0、1,为了避免表述上混淆,将SPSS的分类结果做如下处理: 第1类客户分类编号取为0,即Y=0; 第2类客户分类编号取为1,即Y=1。

首先,尝试全变量法拟合logistic回归模型,SPSS21.0返回结果见表5:

表5 模型系数的综合检验

	卡方	df	Sig.
步骤	4714.288	9	.000
模块	4714.288	9	.000
模型	4714.288	9	.000

这张表是对模型整体显著性的三种似然比检验结果,本文选取显著性水平为0.05,三种检验的Sig.值都接近于0,说明logistic回归模型的系数整体高度显著。

表6 方程中的变量(引入9个指标时)

	B	S.E.	Wald	df	Sig.	Exp(B)
近一年周转率	-437.875	865.488	0.256	1	0.613	0
近一年平均持股时间	899.204	1425.187	0.398	1	0.528	.000
近一年日均仓位	345.017	544.89	0.401	1	0.527	6.90E+149
近一年持股分散度	181.218	339.123	0.286	1	0.593	5.04E+78
近一年最长连续无交易时间	684.337	1084.075	0.398	1	0.528	1.60E+297
近一年普通账户交易量(万)	-230.643	445.808	0.268	1	0.605	.000
近一年最大上涨率	290.567	458.49	0.402	1	0.526	1.55E+126
近一年最大下跌率	370.078	585.071	0.4	1	0.527	5.28E+160
近一年投资收益(元)	4.537	586.101	0	1	0.994	93.402
常量	-370.696	635.333	0.34	1	0.56	.000

这张表是logistic回归的建模结果,也是回归系数的显著性检验结果和变量筛选依据。B表示方程的回归系数,S.E.是其标准误差,Wald是回归系数检验统计量的观察值,Wald统计量形如:

$$\text{Wald}=\left(\frac{B}{S.E.}\right)^2=\left(\frac{\beta_j}{\sqrt{D(\beta_j)}}\right) \quad (17)$$

df表示Wald统计量抽样分布的自由度,Sig.值表示回归系数的检验P值。由此可见,所有偏回归系数P值都大于0.5,高度不显著; 但模型整体高度显著,有充分的理由相信,模型存在严重的多重共线性。

下面通过SPSS的6种逐步回归方法: 前进法(条件)、前进法(似然比)、前进法(Wald)、后退法(条件)、后退法(似然比)、后退法(Wald),进行变量筛选和建模尝试。对于模型整体显著性的检验,考察表“模型系数的综合检验”; 对于参数显著性

的检验，考察“方程中的变量”、“不在方程中的变量”等多张表格，logistic 建模结果也在这里得到。限于篇幅，这里直接展示逐步回归的探究结果。

6 种 logistic 逐步回归的尝试性建模得到了一致的结论，说明本文对某证券公司客户交易数据建立的 logistic 回归模型是稳健的，即选择 7 个解释变量，分别是：近一年周转率、近一年平均持股时间、近一年日均仓位、近一年持股分散度、近一年最长连续无交易时间、近一年最大上涨率、近一年最大下跌率建立模型。

（三）logistic 模型的建立和分析

按照全变量法就 7 个解释变量进行二元 logistic 建模，主要返回结果如下：

表 7　分类表 a

		已观测		
已预测		流失状态分类		百分比校正
		0	1	
流失状态分类	0	2096	0	100
分类	1	1404	0	0
总计百分比			59.9	

注：a 表示模型中包括常量，模型切割值为 0.5.

表 7 说明在没有任何解释变量以前，预测所有的样本点都是正常投资的正确率为 59.9%。

表 8　模型系数的综合检验

	卡方	df	Sig.
步骤	4642.355	7	.000
模块	4642.355	7	.000
模型	4642.355	7	.000

表 8 显示模型整体显著性的似然比检验结果均是：Sig. 值接近于 0，在 0.05 的显著性水平下，有充分的把握拒绝系数全部为 0 的零假设，说明建立的 logistic 模型整体高度显著。

表 9　模型汇总

-2 对数似然值	Cox & Snell R 方	Nagelkerke R 方
71.952a	0.735	0.993

表 9 是对模型整体的拟合优度检验。-2 对数似然函数值为 71.952，较小，说明模型的拟合优度不错[12]。Cox & Snell R 方、Nagelkerke R 方是两个伪决定系数，反映因变量的变化有多大比例可以由自变量解释。因为估计的方法不同，两个伪决定系数的大小通常也不同。Cox & Snell R 方的拟合优度结果为 73.5%，处于 70% ~80% 之间，这是 logistic 回归模型的正常拟合优度范围；Nagelkerke R 方的拟合优度高达 99.3%，相当程度上说明 7 个解释变量对因变量的联合影响几乎完全决定了因变量的变化。

表 10　分类表 a

已观测		已预测		
		流失状态分类		分比校正
		0	1	
流失状态	0	2092	4	99.8
分类	1	2	1402	99.9
总计百分比			99.8	

注：a 表示模型中包括常量，模型切割值为 0.5.

表 10 表明使用该 logistic 回归方程对样本点进行分类，其预测精确度为 99.8%；相比建模前的 59.9%，获得了质的提高，3 500 个样本点中仅出现 6 个误判。实际正常投资的 2 096 位客户中有 2 092位被预测出来，有 4 位错判，正确率达到 99.81%；实际流失的 1 404 位客户中有 1402 位被预测出来，有 2 位错判，正确率达到 99.86%，说明模型有很理想的预测结果，可以为证券公司判断客户流失提供充足的依据，为实践提供较好的参考。

表 11　方程中的变量（引入 7 个指标时）

	B	S.E.	Wald	df	Sig.	Exp（B）
近一年周转率	-27.985	4.511	38.48	1	.000	0
近一年平均持股时间	42.45	6.651	40.733	1	.000	2728080607854.00377500
近一年日均仓位	16.478	2.576	40.925	1	.000	14330543.71
近一年持股分散度	7.717	1.232	39.268	1	.000	2246.951
近一年最长连续无交易时间	33.411	5.219	40.986	1	.000	323783346653858.3
近一年最大上涨率	14.774	2.373	38.756	1	.000	2608494.408
近一年最大下跌率	17.346	2.755	39.656	1	.000	34149564.9
常量	-18.307	2.935	38.916	1	.000	0

由表11可知，所有解释变量的P值均接近于0，说明所有偏回归系数均高度显著，结合模型的整体显著性，综上，用7个解释变量建立的二元logistic回归模型是显著有效的。

模型的函数解析形式如下：

$$\ln\frac{P}{1-P}=-18.307-27.985X_1+42.450X_2+16.478X_3+7.717X_4+33.411X_5+14.774X_7+17.346X_8 \quad (18)$$

其中，X_1——近一年周转率，X_2——近一年平均持股时间，X_3——近一年日均仓位，X_4——近一年持股分散度，X_5——近一年最长连续无交易时间，X_7——近一年最大上涨率，X_8——近一年最大下跌率。

模型的概率形式如下：$P(Y=1|X)=\frac{e^{-18.307-27.985X_1+42.450X_2+16.478X_3+7.717X_4+33.411X_5+14.774X_7+17.346X_8}}{1+e^{-18.307-27.985X_1+42.450X_2+16.478X_3+7.717X_4+33.411X_5+14.774X_7+17.346X_8}}$ (19)

$$P(Y=0|X)=\frac{1}{1+e^{-18.307-27.985X_1+42.450X_2+16.478X_3+7.717X_4+33.411X_5+14.774X_7+17.346X_8}} \quad (20)$$

这里同时给出两种logistic回归的建模结果，因为函数解析形式有利于模型的解释并分析实际结论，概率形式有利于样本数据的代入和对模型的理解。由于P的取值在(0,1)上，而因变量是0-1“开关变量”，与SPSS软件相一致，当$P<0.5$时，将概率四舍五入为0，即返回因变量值0；当$P>0.5$时，将概率四舍五入为1，即返回因变量值1。鉴于此，证券公司在决定是否要对客户采取挽留措施时，可以先将客户交易数据的各项指标代入式(19)，计算客户的流失概率。

进一步，为检验模型的泛化能力，采用事先准备的样本容量为655的测试样本对模型进行验证。泛化能力是指模型对于非样本集的输入，也能给出较精确的输出结果，这是模型有效性和实用价值的重要考量。测试样本的分类结果见表12。

表12 模型对测试样本的分类

已观测		已预测		
		流失状态分类		分比校正
		0	1	
流失状态	0	434	4	99.1%
分类	1	1	216	99.5%
总计百分比			99.2%	

由表12可知，用本文建立的模型预测样本外的客户流失状态，预测精度高达99.2%，655个样本点中仅有5个出现误判。实际正常投资的438位客户中有434位被预测出来，有4位错判，正确率达到99.1%；实际流失217位客户中有216位被预测出来，有1位错判，正确率达到99.5%，说明模型有很高的预测精度和强大的泛化能力，进一步说明了模型具有实际指导意义。下面是对模型基于证券公司业务特点的分析。

近一年周转率高，说明该客户拥有很强的参与度，对证券公司的业务也倾注了大量的时间、精力，这样的客户自然是忠实的，该指标的偏回归系数为-27.985，绝对值较大，说明其与客户流失概率的负相关性很强，是衡量客户流失风险的重要指标。近一年平均持股时间、近一年最长连续无交易时间越长说明该客户的交易活跃度越低，流失风险越大。两指标的偏回归系数分别为42.45和33.411，是最大的两个偏回归系数，表示这两个指标的增加会使客户流失概率上升得非常快，是证券公司需要特别关注的。

观察可见，近一年周转率、近一年平均持股时间、近一年最长连续无交易时间三个解释变量的偏回归系数绝对值都在27以上，其他4个偏回归系数绝对值都在18以下；结合表1可知，反映客户交易活跃度的指标对客户流失概率产生较主要的影响。

近一年最大上涨率和近一年最大下跌率是相对应的两个指标，可放在一起讨论。两个指标的偏回归系数分别为14.774、17.346，都大于0，说明这两个指标与客户流失概率正相关。对这两项指标中至少一项较高的客户进行分类讨论，如下：

类别1(风险投资者)：两指标均较高，从一个侧面说明了该客户愿意进行“高风险，高收益”的投资。该类客户会根据市场行情随机而动，见风使舵，自然不是证券公司稳定的客户群体。

类别2(投资行家)：近一年最大上涨率较高且近一年最大下跌率较低，说明该类客户拥有准确的知觉，对政府政策和市场行情的脉搏有清楚地把握，是证券投资的获利者。同时，因为其头脑的冷静和理性，在一个阶段的投资获益后，明白“股市如赌场”，便倾向于及时撤出资金。该类投资者对行业的认识和经验平均不低于证券公司的普通管理者，所以是公司较难控制的。

类别 3（投资失利者）：近一年最大下跌率较高且近一年最大上涨率较低，从一个侧面说明该客户的择股能力欠缺，可能是证券投资的失利者。该类客户因为前一年的投资遭受损失，可能心灰意冷，从此逐渐退出证券投资领域；也可能会心生换一家公司“卷土重来”的想法。两种情况都会导致该类客户倾向于流失。

因此，对于近一年最大上涨率或近一年最大下跌率较高的客户，证券公司应该多加关注。至于是否采取挽留措施，对于类别 1 和类别 2 的客户只能酌情，因为投入的代价和风险也是证券公司应当考虑的，这需要公司进行商讨和权衡；对于类别 3，证券公司应当努力挽留，采取一定的个性化服务，给予客户投资一些帮助和指导，对政策的颁布和市场行情的变化及时提醒客户，通过人性化的服务与客户建立“人情纽带”，从而避免客户流失。

近一年日均仓位较高，说明客户将较大比例的投资资金用于本证券公司的投资业务，按照直观的理解，这样的客户应该是比较稳定和忠实的；然而，16.478 的偏回归系数说明，近一年日均仓位的提高会显著地增加客户的流失概率，越高的日均仓位预示着越高的流失风险，这说明仅凭经验进行客户流失的判断也是容易导致错误的。

近一年持股分散度越高，在总投资资金一定的情况下，每支股票的投入资金就越少；一位客户越遵循“不要把鸡蛋放在一个篮子里”的原则，该客户就越可能是散户投资者或入门投资者，这样的投资者就越可能缺乏投资主见，出现“跟风”投资的现象，其流失可能性也就越高。该指标的偏回归系数为 7.717，在 7 个指标中绝对值最小，远低于反映客户交易活跃度的 3 个指标，说明它对衡量客户流失风险较次要，证券公司只需有所关注即可。

四、主要结论、建议及展望

（一）主要研究结论

结论一：针对不同的行业、数据、指标应该选择相应最好的模型去预测，经典的数据挖掘模型同样可以取得很好的预测效果。本文采用 logistic 回归预警模型研究证券公司的客户流失问题，取得了很高的预测精度和很强的算法延展性；同时，经典模型保证了结果良好的可解释性和严格的理论基础，避免了过于繁杂的建模和计算。

结论二：对客户流失概率影响显著的 7 项指标中，仅近一年周转率与流失概率负相关，其他 6 项指标均与流失概率正相关，不同的指标对客户流失概率的影响程度有较大差别。反映客户交易活跃度的指标是证券公司实施客户流失预警的关键，包括：近一年周转率、近一年平均持股时间、近一年最长连续无交易时间。

对于近一年最大上涨率、近一年最大下跌率有至少一项数据较高的，证券公司应主要挽留客户类别 3（投资失利者）；对于类别 1（风险投资者）和类别 2（投资行家）的客户，应该在采取挽留措施前有所权衡，避免无谓地争取。近一年日均仓位的提高会显著地增加客户的流失概率，与经验判断不符，需要特别注意。而近一年持股分散度对客户流失的影响较小。

（二）对策与建议

1. 由结论一可知，本文采用 logistic 回归模型取得了非常理想的预测结果。近几年来，客户流失预警领域的建模研究几乎都集中在三个方向：一是组合分类器，二是对经典模型建立改进模型，三是独辟蹊径，采用以随机森林为代表的新兴方法。本文证实了单一的基于传统统计学方法的模型在今天仍然有理论和实际的巨大价值，理想的预测精度和算法的延展性并非总需要建立尽可能复杂的模型。不仅如此，因为传统模型在经济性和可解释性上是后续模型，如：神经网络、支持向量机、组合分类器等，不可比拟的，所以，后者没有显著优势就应该采用前者。这一点是企业和研究人员特别应该了解的。

2. 在本文研究的 9 个客户交易指标中，近一年投资收益被证明对客户流失状态几乎没有影响，近一年最大上涨率、近一年日均仓位越高，客户越倾向于流失，否定了我们的常识。所以，建议证券公司的相关人员在判断流失客户时要基于客户数据建立客观的评价标准，不能仅凭部分从业人员的经验妄下断言，否则很容易对管理人员的政策制定和调整造成误导，以致错失挽留客户的机会。

3. 针对流失风险较高的客户，证券公司应该根据客户价值的高低采取不同的措施。如果是大投资者，像机构投资者、能够为公司带来较多利润的职业股民、资金雄厚的散户投资者，证券公司应当谋求从战略上树立公司和客户双赢的新型客户关系，从战术上给予足够关怀、为他们

制定个性化服务,比如:大势行情的及时传达和分析,政策法规动向的传达和解读,主动为其推荐合适的理财产品,对于极端重要的客户为其量身设计理财产品;如果是普通客户,公司应当考虑挽留成本,见机行事,重在培养客户对公司的情感,如:保证每一位客户享受到较高质量的基本服务,定期采用随机抽样方式对客户进行电话回访,适时的举行活动赠送礼品回馈老客户等。

(三)未来研究方向

1. 在本文研究的基础上加入反映客户基本属性、客户服务情况及交易系统运行质量情况[6]的指标参与分析和建模,这样得到的客户流失预警模型将使证券公司对客户流失状态获得全面的把握。当指标的数量足够多后,如果仍将客户流失状态二分划,提倡采用支持向量机参与建模;如果将客户流失状态做多分划,提倡建立组合分类器,并保证多分类 logistic 回归模型作为子分类器参与建模。

2. 客户流失预警建模的前期工作可以与客户细分、客户价值的综合评价相结合,这样能为证券公司建立客户流失预警系统提供更科学的建议。因为毕竟不同客户对于公司的价值差异极大,并非只有将要流失的客户才值得公司挽留,当前稳定投资的客户,公司就可疏于关注和关心。

参考文献:

[1] 仲 继. 电信企业客户流失预测模型研究[D]. 西安科技大学硕士学位论文,2014.

[2] 姜晓娟,郭一娜. 基于改进聚类的电信客户流失预测分析[J]. 太原理工大学学报,2014,44(4):532-536.

[3] 王建仁,李 妮,段刚龙. 基于信息融合的电信客户流失预测研究[J]. 计算机工程与应用,2015,52(1):71-76.

[4] 杨孝成. 基于决策树的移动通信用户流失预警模型研究与实现[D]. 中国海洋大学硕士学位论文,2014.

[5] 王 卉. 基于服务修复理论的证券公司客户流失问题[D]. 江西财经大学硕士学位论文,2009.

[6] 杜修平,王 中. 基于决策树的证券客户流失模型[J]. 计算机应用与软件,2009,26(9):230-233.

[7] 吴 斌. 基于 Logistics 回归算法的证券客户流失预测模型及应用[J]. 金融电子化,2013,11(7):65-67.

[8] 于彩娴,赵治荣. 银行客户流失预测的数学建模分析[J]. 长春工业大学学报(自然科学版),2013,34(1):5-8.

[9] Verbeke W,Dejaeger K,Martens D,et al. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach[J]. European Journal of Operational Research,2012,218(1):211-229.

[10] 柳 婷. 基于数据挖掘的银行客户流失模型分析研究[D]. 重庆大学硕士学位论文,2008.

[11] 陈彦光. 人口与资源预测中 Logistic 模型承载量参数的自回归估计[J]. 自然资源学报,2009,24(6):1105-1114.

[12] 梁 锋. 数据挖掘技术在寿险客户流失中的应用[J]. 电子科学技术,2015,4(1):104-107.

责任编校:陈 强,王彩红

Customer Churn Warning Analysis on Securities Companies Based on Logistic Model

ZHENG Yu-chen, LV Wang-yong

(*School of Mathematics and Software Sciences, Sichuan Normal University, Chengdu 610011 China*)

Abstract: With the rapid development of China's economy and the deepening of economic globalization, customer churn has become more important than grabbing customers for securities companies. Starting from the index reflecting the customer transactions, K-means cluster is used for obtaining customer churn state. Through 6 kinds of stepwise regression method to variable selection, a logistic customer churn warning model is set up in this paper. Moreover, the generalization ability of the model is tested and analysis based on business characteristics of securities companies´are given. The results show that: the customers´trading activity index is the key to the implementation of customer churn warning of the securities companies. Furthermore, an effective model and feasible suggestion are put forward to targeted customer retention for the securities companies.

Key words: customer churn warning model; logistic regression; data mining; securities companies