DOI:10.19473/j.cnki.1008-4940.2018.02.006

商业银行贵宾客户流失预测研究

卢美琴[1]，吴传威[2]

（1. 福建商学院 国际经济与贸易系，福建　福州，350012；
2. 中国农业银行福建省分行 科技与产品管理部，福建　福州，350003）

[摘　要]　随着我国银行同业的竞争加剧以及互联网金融的冲击，商业银行贵宾客户流失现象越来越严重。从海量的客户信息中挖掘出对流失有重要影响的因素，从而建立有效的银行客户流失预警体系显得尤为必要。结合我国商业银行业务现状，综合运用属性相关性、决策树分析等方法，创建商业银行贵宾客户流失预警模型，同时利用聚类分析对流失客户进行群体细分，针对每个群体给出其特征描述和挽回措施。

[关键词]　决策树；聚类分析；贵宾客户；客户流失

[中图分类号]　F832.2　　[文献标识码]　A　　[文章编号]　2096-3300（2018）02-0031-06

一、引言

在经济转型大背景下，我国商业银行的经营形势面临着天翻地覆的变化。金融脱媒和利率市场化进程逐步加快，银行利差大幅缩窄[1]，银行间的竞争更加激烈，互联网金融企业开始抢占商业银行的传统领域，侵蚀银行利润空间。近年来银行对公业务已经成为红海战场，越来越多的商业银行将经营重心从对公业务向个人业务转移，个人零售客户成为竞争焦点。向零售业务转型升级已经成为近年来银行业应对互联网金融发展、经济新常态以及监管趋严态势的必然选择。

根据二八定律，20%的客户贡献了 80%的利润。统计分析表明贵宾客户是商业银行的主要个人客户群体，该群体的扩展及维系对银行的经营起到至关重要的作用，成为商业银行日常经营的重中之重。然而，随着客户金融消费需求升级，客户对金融服务的要求进一步提高，金融市场供求格局也随之发生变化，多种因素共同作用下，银行贵宾客户群体的不稳定性增加。客户流失不仅会增加银行的营销费用和机会成本，还会对银行声誉产生负面影响[2]。研究表明，对银行业而言，客户流失对利润有着巨大的影响，客户流失率减少 5%，能给企业带来 30% ~85% 的利润增长。发展新客户的成本是挽留客户的 5 ~7 倍，而挽留客户的成功率却是发展新客户成功率的 16 倍[3]。因此，识别影响客户流失的关键因素，有效预测客户流失可能性并制定相应的挽回措施，防止客户流失，是商业银行提升核心竞争力的关键。

收稿日期：2018-02-26

作者简介：卢美琴（1985-），女，福建泉州人，讲师，硕士，研究方向：数据挖掘、网络营销；
吴传威（1985-），男，福建宁德人，工程师，硕士，研究方向：银行经营数据分析。

国内外学者也对此进行了大量的研究，包括流失原因研究、流失预测研究和客户挽留机制研究，主要应用神经网络、决策树、贝叶斯网络、支持向量机等模型。如梁礼明等[4]使用 BP 神经网络对客户流失进行预测；王未卿等[2]对客户流失产生重要影响的预测变量进行分析，并通过建立 Cox 比例风险模型，对客户流失的可能性进行预测；Prasad 和 Madhavi[5]分别用 CART 和 C5.0 两种分类技术研究了商业银行客户流失行为；贺本岚[6]对支持向量机和 Logistic 回归模型在银行客户流失预测的效果进行了对比。通过对几种方法的对比发现，模型各有优缺点：贝叶斯便于先验知识和样本数据的结合，但是如何取得先验知识是个难题[7]；神经网络精度高，但其规则解释性差；支持向量机 SVM 分类正确率高，但其求解需要较大量的计算，对于实际商业环境的大数据来说对资源要求太高。相比较而言，决策树分类算法以其计算量小、规则解释性强等特性，特别适合商业银行开展大量客户的流失预测分析。从已有银行客户分析研究可以看出，现有研究主要集中在流失预测准确性的提高，缺乏针对贵宾客户群体的流失研究，并且对定位出的流失客户的流失挽回环节研究涉及较少。因此，针对贵宾客户建立流失预测模型，强化流失预测环节与流失挽回环节的关联，对提高银行客户流失挽回工作效率、降低客户流失率有显著作用。

综上所述，保留老客户、防止贵宾客户流失对于商业银行的经营稳定具有重要意义。而防止客户流失的关键在于能够提前定位可能流失的客户，采取挽留措施，降低其流失意愿。本文以某商业银行某分行为例，具体探讨如何利用决策树方法建立贵宾客户流失预测模型，并利用聚类分析方法对流失客户进行细分，针对每个群体给出其特征描述和挽回措施。

二、理论基础

(一) 决策树原理

目前，国内外客户流失预测算法使用最为广泛的是回归、决策树和人工神经网络。而其中决策树由于其良好的规则解释能力和学习效率，成为广泛采用的预测算法。决策树（Decision Tree）运用概率方法对决策中的不同方案进行比较，从而得出最优方案，由于这种决策分支画成图形很像一棵树的枝干，故称决策树。其具体算法如下[8]：

设 D 为一个包含 $|D|$ 个数据样本的集合，类别属性有 m 个不同的值，对应于 m 个不同的类别集合 C_i，$i \in \{1, 2, 3\cdots m\}$，$|C_i|$ 是类别集合 C_i 中的样本个数，对 D 中的元组分类所需的期望信息为：

$$I(D) = \sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

其中，$P_i = |C_i|/|D|$ 表示一个数据对象属于类别 C_i 的概率。

假设按照属性 A（取值为 $\{a_1, a_1\cdots a_v\}$）将 D 划分成 v 个不同的类 $\{D_1, D_1\cdots D_v\}$，那么使用属性 A 对当前样本集进行划分的信息熵为：

$$I_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{D} I(D_j) \quad (2)$$

信息熵 $I_A(D)$ 的值越小，表示利用属性 A 进行子集划分的结果越好。

这样，利用属性 A 对当前分支节点进行相应子集划分所获得的信息增益为：

$$Gain(A) = I(D) - I_A(D)$$

C4.5 算法为了避免结果倾向于具有大量值的属性，将信息增益定义为：

$$GainRate(A) = \frac{Gain(A)}{SplitI_A(D)} \quad (3)$$

其中：$$SpligI_A(D) = -\sum_{j=1}^{v} \left|\frac{D_j}{D}\right| \times \log_2 \times \left(\frac{D_j}{D}\right)$$

在每个分支节点上，C4.5 算法计算每个属性的信息增益率，从中选择信息增益率最大的属性作为

在该节点上进行子集划分的属性，直到信息增益率低于某一特定阈值时停止决策树的构造。C5.0 是 C4.5 的升级版，在执行效率和内存使用等方面都进行了改进，特别适合于大数据集上[9]。

（二）客户细分原理

客户细分主要指根据客户的价值、需求和偏好等综合因素对客户进行分类，分属于同一客户群的消费者具备较高的相似性，而不同的客户群间存在明显的差异性。通过客户细分，企业可以更好地识别不同客户群体对企业的价值及其需求。

在数据挖掘中，往往通过聚类的方法来实现细分。K-Means 算法是一种经典的聚类算法，对处理海量数据有着较高的伸缩性，且效率较高，因此特别适用于银行客户的细分。指定聚类簇数 K，算法随机选取 K 个记录作为初始中心，分别计算每个记录到 K 个中心的的距离，按距离最近原则将每个记录都归属到 K 个簇；按平均值方法计算每个簇的中心，再次计算每个记录到 K 个中心的距离，重新调整每个记录的归属……，直至满足设定的循环次数或簇归属稳定。

其基本函数为：

$$\forall p \in PC,\ distance(p,\ getCluster(p)) \leq \min_{0<i\leq M}\{distance(p,\ C_i)\} \quad (4)$$

其中，p 表示样本，PC 表示样本集合，distance( ) 表示样本与聚簇中心的距离，getCluster 表示样本所属聚簇中心，M 表示聚簇个数，表 C_i 示第 i 个聚簇[10]。

三、实证分析

（一）数据来源

根据研究目的，本文选取观察期内资产下降 90%以上的客户作为客户流失定义进行分析，具体客户流失定义：客户前三个月（$T-2$，$T-1$，T）月日均资产有 10 万以上且在年日均资产 50%以上，随后三个月（$T+1$，$T+2$，$T+3$）月日均资产流失达 90%以上，且未来三个月（$T+4$，$T+5$，$T+6$）未恢复。

数据集来源于某商业银行数据仓库，选取的时间窗口为 2016 年 9 月到 2017 年 8 月，经过数据清洗与处理，共得到 2 758 289 条资料完整的客户记录，其中流失客户数为 71 011 个，流失客户占比为 2.57%。该数据集为典型的不平衡数据集，为了减小流失客户与非流失客户之间的比例差距，提高模型对流失客户的识别能力，通过随机欠抽样法，即减少多数类样本数量，构造新数据集，最后选取 142 022 条记录，其中流失客户与非流失客户各占比 50%。然后按照 2：1 左右的比例划分训练样本集和验证样本集，分别用于训练模型和验证模型有效性。

（二）基于决策树方法的流失预测模型构建

1. 预测指标筛选

预测指标对于决策树模型以及试验结果具有重要意义，指标选取将最终影响模型预测的有效性。参考以往研究并结合该行实际业务情况，选取了 50 个初始指标。而这些指标是否对客户流失产生影响需要进行相应检验，并且这些指标间可能存在重复信息需要排除。因此对初始指标进行约简，主要步骤为：首先，检验每个属性指标对客户是否流失的影响程度，剔除相关系数小于 0.7 的指标；其次，按每个属性对客户是否流失的相关性由大到小排序，将其他属性与当前属性进行相关性分析，将相关性大的属性删除，以此来消除冗余。

使用 Pearson 相关系数检验来检验 2 个变量之间的相关性，其值越接近 1 则表明正相关性越大，其值越接近-1 表明负相关性越大，其值越接近 0 则表明相关性越小。通过对初始的 50 个指标进行相关性分析后，确定出与客户流失关联性较大的 15 个特征用于构建决策树模型，如表 1。

表 1 客户流失相关因素

Tab. 1 Factors for customer loss

属性类别	变量名称	属性类别	变量名称
自然属性变量	年龄	行为特征属性变量	未来三个月到期整整占总资产超 50%
	性别		整整销户
账户及产品变量	是否有理财		活期销户
	是否有贷款		大小额转出笔数超 8 笔
	有效活期账户数		贷款提前结清
行为特征属性变量	资金流出占总资产超 50%		贷款到期结清
	大小额跨行转出占总资产超 30%		未来三个月到期理财占总资产 80%

2. 预测误判代价矩阵的确定

决策树 C5. 0 算法的一个显著改进在于引入了代价矩阵，可以有效地减小误判的代价。在实际对客户进行流失判断的过程中，可能会将非流失客户误判为流失客户或将流失客户误判为非流失客户，对于商业银行来说，前者可能仅仅是客户维护人员打一个电话的花费，后者则可能损失一个重要客户，使银行蒙受较大损失。相比较而言，后者给商业银行带来的损失要远大于前者。通过与个人金融部等贵宾客户主管部门核算，确定代价矩阵，见表 2。误判代价矩阵表明，将实际会流失客户标识为非流失的代价，是将实际非流失客户标识为流失客户代价的 10 倍。

表 2 代价矩阵

Tab. 2 Cost matrix

流失类型	预测流失	预测非流失
实际流失	0	10
实际非流失	1	0

（三）预测效果分析

使用训练样本集训练生成决策树模型，使用验证样本集对模型预测的稳定性进行考察。为了验证决策树模型的预测效果，引入业界普遍使用的两个评价模型有效性的指标:

流失覆盖率 = 正确预测流失客户数/总流失客户数

预测准确率 = 正确预测流失客户数/总预测流失客户数

流失覆盖率反映的是模型最终查找出的真实流失客户占实际总流失客户的百分比; 预测准确率反映的是模型标记出的流失客户中真正流失的百分比。从模型对训练集和验证集的预测结果来看（见表 3），预测模型能够查找出 61%左右的流失客户，且预测准确率超过 82%，具备较强的实用性。

表 3 预测效果

Tab. 3 Prediction effect

数据集	流失覆盖率%	预测准确率%
训练集	61. 23	82. 91
测试集	60. 79	82. 89

（四）流失客户细分与挽回措施

流失客户中，由于客户年龄、资产结构、交易习惯等的差异，其流失原因和流失特征也各不相同，如果采用相同的挽回策略，难以起到针对性营销的效果。对流失客户进行细分，根据其不同特征划分为不同的流失群体，针对每个流失群体进行分析，描述其群体特征，并给出相应的挽回措施，将有助于提高客户维护人员的工作效率和效果。因此，利用数据挖掘技术对流失客户进行聚类细分，对每一个细分群体分别进行群体特征分析，见表 4。

表 4　聚类因素

Tab. 4　Factors for clustering

维度	变量	解释
自然属性	年龄	客户所处年龄段对行为有很大影响
财务能力	最高月均资产	预示客户的潜在财力
财务能力	年均资产	财力的主要指证
交易习惯	半年交易次数	表明客户对银行产品服务的使用程度
品牌忠诚	最大活期账户年龄	表明对银行的忠诚度

参考现有银行客户聚类分析常用指标及实际可获得性，获取到表 4 中的聚类指标，包括：自然属性、财务能力、交易习惯、品牌忠诚等。利用 K-Means聚类算法，将流失贵宾客户细分为四个群体，对四个群体进行分析，分别定义族群标签，描述群体特征及提出流失挽回措施，具体结果见表 5。

表 5　客户细分结果

Tab. 5　Results of customer segmentation

族群标签	群体特征	挽回措施
时尚小白领	这是一个年轻的新客户群体（平均年龄 35 岁，最长活期账户年龄 43 月），交易较活跃（半年交易次数 25 次）。交易以网银、ATM 为主，是四个群体中柜台交易占比最低的。	其资产以活期为主，向其推荐短期理财、定活通等产品增强客户粘性。
中产安居客	该群体忠诚度很高高（平均最大活期账户开户时间为 97），对产品服务的使用频度较高（半年各类交易次数达 13），具有一定的理财习惯，主要为定期存款。该群体使用 ATM、网银及柜台交易的比例比较平均，是所有群体中使用柜台交易最频繁的。	重点加强其来到柜台的服务，增强服务满意度。
都市新贵族	该客户群体客户数极少，但平均最大月均资产达到四千多万，交易较活跃（交易次数达 19），较弱的品牌忠诚度（最大活期账户年龄为 32 个月）；该群体资产以理财为主。	需要重点关注的客户，增强其品牌认同感，特别关注其理财资产的到期时间。
青年精英	该客户群体较年轻（44 岁），交易较活跃（半年交易次数 18 次），是银行的忠实客户（最大活期账户年龄为 103）；具有较强的理财习惯；交易以自助渠道及电子渠道为主。	其资产以活期为主，向其推荐短期理财、定活通等产品增强客户粘性。

客户经理根据流失贵宾客户归属的群体特征及挽回措施建议，结合客户资产结构、近期交易特征以及客户未来三个月理财、定期产品到期情况，可以实现根据客户特征进行差异化客户维护。

四、结束语

随着内外部经营形势的变化，个人客户流失已经成为商业银行必须解决的问题之一。本文具体分析了对商业银行经营效益起到至关重要作用的贵宾客户的流失影响因素，构建贵宾客户流失预测模型，可以有效识别潜在流失贵宾客户；同时，利用聚类算法对流失贵宾客户进行细分，针对每一个细分群体进行特征描述和制定挽回策略，可以帮助客户维系部门有效提高客户流失挽回工作的效率和效果，也为商业银行进行贵宾客户流失挽回提供了一个新思路。

参考文献：

[1] 贺本岚．支持向量机模型在银行客户流失预测中的应用研究[J]．金融论坛，2014（9）：70-74.

[2] 王未卿，姚娆，刘澄，等．商业银行客户流失的影响因

素 [J]. 金融论坛，2014 (1)：73-79.

[3] 肖进，刘敦虎，贺昌政. 基于 GMDH 的“一步式”客户流失预测集成建模 [J]. 系统工程理论与实践，2012，32 (4)：808-813.

[4] 梁礼明，翁发禄，丁元春. 神经网络在客户流失模型中的应用研究 [J]. 商业研究，2007 (2)：55-57.

[5] PRASAD D，MADHAVI S. Prediction of churn behavior of bank customer customers using data mining tools [J]. Business Intelligence Journal，2012，5 (1)：96-101.

[6] 贺本岚. 支持向量机模型在银行客户流失预测中的应用研究 [J]. 金融论坛，2014 (9)：70-74.

[7] 洪丽平，覃锡忠，贾振红，等. 基于后验概率支持向量机在客户流失中的预测 [J]. 计算机工程与设计，2016，37 (2)：430-432.

[8] 王红武，朱绍涛，蔡海博. 基于决策树算法的上市公司股东行为研究 [J]. 数理统计与管理，2017，36 (1)：139-150.

[9] 杨胜刚，朱琦，成程. 个人信用评估组合模型的构建——基于决策树—神经网络的研究 [J]. 金融论坛，2013 (2)：57-61.

[10] MUDA Z，YASSIN W，SULAIMAN M N，et a1. Intrusion detection based on K - Means clustering and Naive Bayes classification [C]. California：International Conference on Information Technology in Asia，2011.

A Study on VIP Customer Churn of Commercial Banks

LU Meiqin，WU Chuanwei

(1. Department of International Economy and Trade，Fujian Business University，Fuzhou 350012，China；
2. Department of Science，Technology and Product Management，
Agricultural Bank of China Fujian Branch，Fuzhou 350003，China)

Abstract: With the increasing competition within Chinese banks and the impact of Internet finance，the VIP customer churn of commercial banks is becoming more and more serious. Therefore，it is essential for banks to establish an effective early warning system for customer churn by discovering the influential factors of the churn from the mass of customer information. Based on the present condition of the commercial banks in China，this study creates a warning model of commercial banks' VIP customer churn through an integrated use of attribute correlation and methods of decision tree analysis. Meanwhile，the cluster analysis is used to subdivide the losing customer groups with the description of characteristics and redemptive measures for each group.

Key words: decision tree；cluster analysis；VIP customers；customer churn

(责任编辑：杨成平)