DOI:10.15923/j.cnki.cn22-1382/t.2013.01.002

银行客户流失预测的数学建模分析

于彩娴， 赵治荣*

(太原工业学院 理学系，山西 太原 030008)

摘 要：利用数据分析系统软件，用 4 种比较常用的数学建模方式进行分析，并对分析的结果结合业务需求做了比较，最后选择逻辑回归模型进行了客户流失预测，并对银行客户保留问题提出了有针对性的对策建议。

关键词：数学建模分析；流失预测；数据挖掘；SAS 软件；逻辑回归

中图分类号：O 29 **文献标志码**：A **文章编号**：1674-1374(2013)01-0005-04

Mathematical modeling and analysis on bank customer churn prediction

YU Cai-xian， ZHAO Zhi-rong*

(Department of Science of Taiyuan Institute of Technology, Taiyuan 030008, China)

Abstract: With Statistical Analysis System (SAS) software, we apply four different mathematical models to predict the loss of bank customers, and then compare the results with the real cases. The logistic regression model is chosen to estimate the customer churn, and a proposal is offered for the reservation of bank customers.

Key words: mathematical modeling and analysis; customer churn; data mining; statistical analysis system; logistic regression model.

1 银行客户流失研究问题界定

对于客户流失目前还缺乏统一的定义，通常来说，客户终止与本公司的服务或转向其他公司提供的服务，就被认为是客户流失。客户流失按照是否为客户的主观意愿可以分为自愿流失与非自愿流失。由于流失原因的不同，各种流失的表现相差也比较大，因此，无法找到一种能够预测所有流失的模型和方法，只能针对流失的不同类型分别做相关的分析和研究，建立一个能够相对更加准确预测该类流失的模型。

2 常用客户流失模型及其 SAS 软件实现

客户流失预测有很多种模型，面对不一样的商业需求，各种模型的评价效果各不相同，只有在具体情况下进行比较分析，来选择相对更优的模型。文中采用了 4 种常见的建模方式。

收稿日期：2012-11-09

基金项目：山西省自然科学基金项目(2012011002-2)

作者简介：于彩娴(1976－)，女，汉族，山西太原人，太原工业学院讲师，硕士，主要从事应用数学方向研究，E-mail：hanzheng060626@qq.com. ＊联系人：赵治荣(1960－)，男，汉族，山西太原人，太原工业学院副教授，硕士，主要从事应用数学方向研究，E-mail：hanzheng060626@qq.com.

2.1 4种常用的客户流失模型

2.1.1 逻辑回归模型

在定量分析研究中,线性回归模型是比较流行的统计方法。然而,在现实世界中,线性回归模型对分类输出变量无法做出很好的解释,因此,在设置分类变量时,通常采用的一种方法是对数线性回归模型。当对数线性模型中的二分类变量被当作输出变量并定义为一系列自变量的函数时,对数线性模型就变为逻辑回归模型[1]。逻辑回归模型有很多优点,表现在:可以处理二值因变量;不需要满足其它多变量技术所要求的假设;自动进行变量选择;模型结果清晰,易于对业务部门解释等。

2.1.2 决策树模型

决策树是一种挖掘数据中潜在的分类规则的算法。每个决策或事件都可能引出两个或多个事件,导致不同的结果,把这种决策分支画成图形很像一棵树的枝干,故称决策树。

2.1.3 人工神经网络模型

人工神经网络是通过模仿动物神经网络行为,进行分布式并行信息处理的算法数学模型。它通过调整内部大量节点之间相互连接的关系,达到处理信息的目的。

2.1.4 决策树结合逻辑回归

模型会采用多个变量,但并不是每个变量都会对结果产生显著性影响,每种预测模型都有自己的预测方式剔除对结果影响并不显著的变量,而决策树结合逻辑回归即是用决策树剔除对结果影响不显著的变量后,用逻辑回归进行预测[2]。

2.2 客户流失分析流程

文中使用SAS Enterprise Miner系统实现客户流失建模。客户流失模型逻辑SEMMA分为数据采样(Sample)、数据分析(Explore)、模型调整(Modify)、模型建立(Model)、模型应用评估(Assess)。建立客户流失分析过程的流程如图1所示。

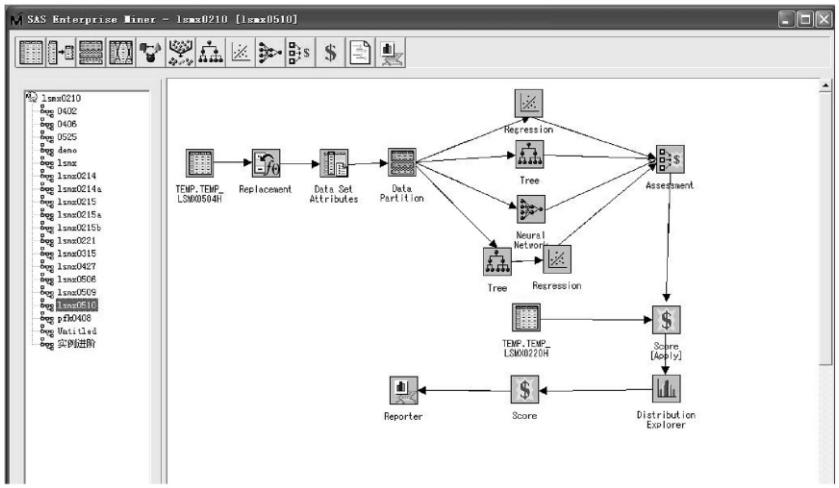

图1 SAS客户流失模型过程介绍

具体流程如下:

1)系统预先设定模型的任务和测量的标准,客户数据从数据输入节点(Input Data)导入,模型用图标的形式展现通过统一计算的变量集合和统计分布结果;

2)替换节点(Replacement)替换数据集中的缺失值或某些非缺失值;

3)数据集属性节点(Data Set Attributes)修改数据集的名称、描述、角色等属性;

4)数据分割节点(Data Partition)将客户数据分为训练数据和预测数据,预测数据用来辨别适合的模型;

5)选用合适的模型建模预测;

6)评估节点(Assessment)来评价4个模型的分析质量;

7)得分节点(Score)对新数据集进行预测;

8)报告节点(Reporter)来生成分析报告。

3 银行客户流失预测实证研究

3.1 数据集描述

文中采用的是某银行信用卡的历史真实数据库提供的客户信息来构建客户流失模型。以该银行为对象,从其数据库中随机抽取约3 500个客户做研究对象。在满足以上定义条件的客户中,

发现该样本数据，流失的客户概率大约是 300 多个，大约占总样本量的 10%。

3.2　客户流失约定

本案例中该银行信用卡客户流失标准确定为：从第一次消费时间距今已满 9 个月；消费次数大于或等于 10 次；客户已经销卡或者连续 6 个月没有任何交易。

3.3　变量选择

考虑信用卡业务的特殊性，其数据准确度高于一般类型的金融业务，且各方面数据非常完善，文中采用了尽量多的变量做全方面分析，然后通过模型再剔除对结果影响较小的变量。

两类变量：

第一类是客户基本信息变量，包括性别、年龄、职务、行业性质、教育程度、年收入、有无他行信用卡、有无关联扣款；

第二类是客户交易变量，包括总消费次数、总消费金额、总取现次数、总取现金额、存活天数、总贡献、累计逾期次数、有效存活天数、平均使用额度、近 1 月交易次数、近 2 月交易次数、近 3 月交易次数、近 4 月交易次数、近 5 月交易次数、近 6 月交易次数、连续降低交易次数的月数[3]。

3.4　实证比较

随机选取 2/3 的数据分为训练数据集，用于模型的建立和调整；1/3 的数据分为验证数据集，用于评估模型的预测效果。然后按照 SAS 的客户流失模型逻辑 SEMMA 来建立整个流程。

按照上面步骤执行完挖掘流程后，SAS 提供了评估节点，可以对 4 种模型生成 Captured Respond图来进行一个比较直观的比较，如图 2 所示。

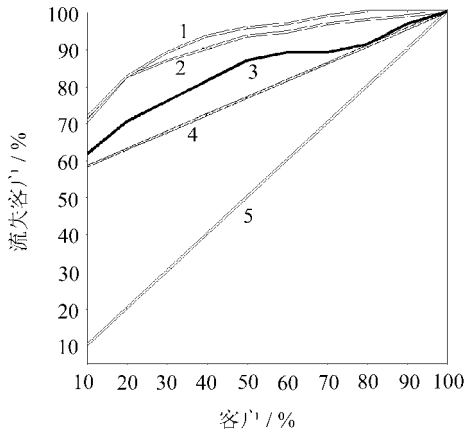

图 2　Captured Respond 图

在图 2 中，横轴表示客户的总百分比，纵轴表示流失客户总数的百分比。斜率为 1 的直线（线 5）代表的是客户的流失率，不使用任何模型时在 10%的客户中只能分辨出 10%的流失客户，在 30%的客户中只能分辨出 30%的流失客户。

另外 4 条折线是基于 4 种建模方式的响应图（线 1：逻辑回归；线 2：神经网络；线 3：决策树结合逻辑回归；线 4：决策树），可以看出，预测效果要大大好于不用任何模型的随机效果。以 2 号折线（从上数第二条）表示的神经网络模型为例。在前 10%的客户中，可以分辨出 70%多一点的流失客户，前 20%的客户可以分辨出 83%左右的流失客户。在前 10%的客户中，神经网络模型的预测效果是随机的，提升性能是 7（70%/10%＝7）。举例说明，如在预测 1 000 个客户的流失情况时，在不采用任何模型随机的情况下，取 100 名客户，只能随机分辨出 10%，即 10 个客户，而采用神经网络模型时，取评分前 100 名客户，可以分辨出 70%，即 70 个客户是流失客户，大大优于随机的结果。

在前 10%的客户中，决策树的提升性能大约是 5.8，逻辑回归是 7.2，决策树结合逻辑回归大约是 6.3。比较效果，逻辑回归＞神经网络＞决策树结合逻辑回归＞决策树。

从图 2 的结果观察，逻辑回归的总预测效果最佳，但如果单独从流失客户的预测率来观察结果略微有不同。查看 SAS 提供的诊断图（Diagnostic Chart），观察逻辑回归和神经网络的实际预测结果如图 3 和图 4 所示。

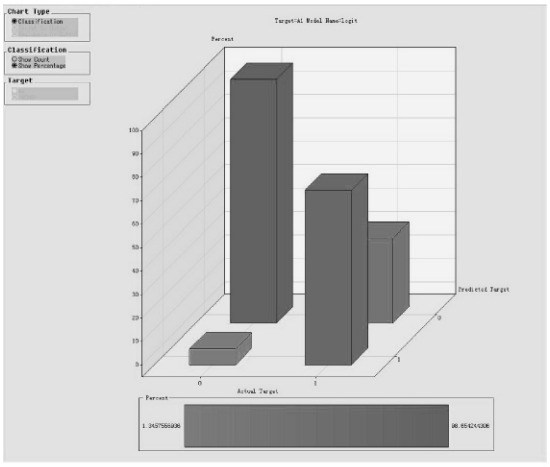

图 3　逻辑回归的诊断图

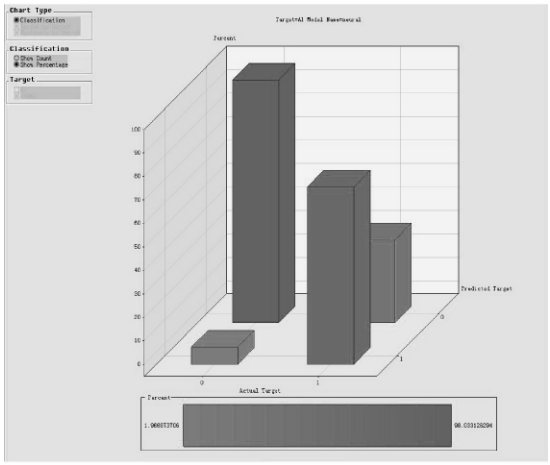

图 4 神经网络的诊断图

图中,右前方的柱体表示将“实际流失”的客户预测为“流失”客户的比例,右后方的柱体表示将“实际流失”的客户预测为“未流失”客户的比例,左前方的柱体表示“实际未流失”的客户预测为“流失”客户的比例,左后方的柱体表示“实际未流失”的客户预测为“未流失”客户的比例。

3.5 采用逻辑回归做银行客户流失预测分析

在本例中,虽然神经网络与逻辑回归在预测效果上差异不大,但逻辑回归能够分析出哪些变量在何种程度上影响了结果,能够给出流失原因的线索,因此,文中采用逻辑回归对该流失问题做为最后采用的模型。

3.6 逻辑回归的结果输出

从结果分析,总消费次数、客户存活天数、客户有效存活天数、近 1 月消费次数、近 1 月取现次数、近 2 月取现次数、近 2 月使用额度比例/平均使用额度、近 3 月使用额度比例/平均使用额度、教育程度、行业性质、是否与借记卡关联扣款、最近一次交易距今的天数是左右信用卡流失的重要因素。

4 客户保留策略

4.1 未流失客户的保持

1)针对单个客户的保持,可根据在模型中所分析出的关键特征设定有流失趋向的临界值,比如设定最近交易距今天数天的临界值为 290 d,最近第 3 月使用额度比例/平均使用额度的临界值 0.75 等关键特征的临界值。通过进一步对这些特征值进行系统的处理,建立客户流失预警系统。并根据预警的结果对客户采取有针对性保持策略,对没有建立关联扣款的客户,鼓励其建立关联扣款关系。

2)对客户群体进行营销,行业性质、教育程度的不同对客户流失有一定影响,为了提高这部分客户的忠诚度,在机关事业单位、股份制企业、私营企业中,针对大专、本科、硕士及以上的客户群体采取与传统的营销手段相结合的客户保持策略[4]。

4.2 预测为流失的客户的挽留

结合流失客户的流失细分和客户的贡献值采取不同的客户挽留策略:

1)预测为高流失风险的客户,可放弃这部分客户;

2)预测为中流失风险的客户,可开展更多的营销服务,提供延长客户生命周期的服务;

3)预测为低流失风险的客户,可提供成本相对较低的针对性营销,也能达到延长客户生命周期的目的[5]。

5 结 语

不同的商业需求会有不同的建模方法来解决,但无论什么方法都是服务于商业需求的。一个真正的商业模型是一个团队在一个较长的周期内,持续地和市场人员进行沟通,不断调整模型和参数,最终建立一个能够适应市场环境变化的动态的模型参数系统。

参考文献:

[1] 简毅明. 数据挖掘在信用卡客户流失分析中的应用[J]. 数据库技术,2007(1):55-59.

[2] 王文硕. 我国商业银行信用卡客户流失实证研究[J]. 金融理论与实践,2008(11):17-21.

[3] 姚志勇. SAS 编程与数据挖掘商业案例[M]. 北京:机械工业出版社,2010:128-133.

[4] 钱苏丽. 基于改进支持向量机的电信客户流失预测模型[J]. 管理科学,2007(2):54-58.

[5] Medova E A. Measuring risk by extreme values[J]. Risk,2000(11):20-26.