DOI:10.19789/j.1004-9398.2019.05.004

基于决策树的存量客户流失预警模型*

杨 荣 赵娟娟 贾郭军

(山西师范大学数学与计算机科学学院,山西 临汾 044600)

摘要: 本研究基于价值分析角度,结合用户套餐资费情况,使用CART算法建立决策树模型,并采用交叉验证法来选取最优决策树,并与逻辑回归算法相比较,得到了较好效果,同时对流失用户属性进一步探索,对运营商精准启动预警挽留和维系策略提供决策支持.

关键词: 存量客户; 决策树; CART算法; 预警挽留

中图分类号: TP399

0 引 言

随着全业务运营,各运营商之间的竞争日趋激烈,电信行业的传统业务面临巨大的压力. 因此,各运营商都加大了在存量市场的争夺,对于不同业务的后进入者,市场竞争往往选择跟随策略,竞争目标多为存量市场,导致客户在各运营商之间的流转加快,各运营商拆机率均居高不下. 与此同时,存量流失造成用户发展效率低,资源消耗严重,损害了企业价值,存量下滑成为公司发展的重要风险. 因此,实现存量用户基本稳定、提升客户忠诚度,提升客户价值已成为各大运营商的经营理念之一. 有效降低存量客户的流失率才能固其根本.

为加强针对存量经营工作的大数据支撑能力,形成以客户为中心的管理运营模式,就需以存量客户为切入点,并建立存量客户流失预警模型,通过加强对高流失概率的用户进行提前预警维系,以达到存量客户保有率的提升效果.

1 国内外研究现状

客户流失预警模型的应用研究如今已非常广泛,常用算法包含神经网络、决策树、随机森林等. 李爱民[1]采用K-means聚类分析和Logistic回归建立客户流失预警模型相结合的算法来分析各种因素对客户流失的影响程度. 周静等[2]构建了计量经济模型,并研究了公司保留策略与延长客户生命周期之间的相关性; 陈纪铭[3]使用朴素贝叶斯算法建立了学员流失预警模型,但该模型假设属性之间相互独立,但在实际使用中属性个数比较多或者属性之间相关性较大时效果较差. 林明辉[4]利用神经网络的自适应算法,将代表离网用户行为特征的45个指标进行样本训练,最终得到客户流失行为倾向的判断模型; Huang和He[5]提出了优化PSO算法和BP神经网络算法相结合的方法来建立企业客户流失预警模型,但神经网络模型需要大量的参数,且黑盒操作,不能观察之间的学习过程,输出结果难以解释,会影响到结果的可信度和可接受程度; 同时大多数公司都使用建模工具来直接建立流失预警机制,准确率仍然较低,本文使用python环境进行建模,并使用决策树模型,并对模型进行优化,使模型算法结果的预测准确率更高.

2 相关算法介绍及算法实现

2.1 CART算法概述

本文使用的是决策树中的CART算法,决策树学习一般分为三个步骤,即特征的选择、决策树的建立、决策树的剪枝; 该算法是一种二叉树形式的决策树算法,其中,二叉树算法只把每个非叶节点引申为两个分支,首先进行二元分割,将样本数据划分成两个子集,其次对子集再分割,自顶向下不断递归生成树,直至分支差异结果不再显著下降,

收稿日期: 2018-11-07

* 山西师范大学校级研究生创新项目(01053002)

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

即分支没有意义,决策树建立完成[6]. 因此确定分枝标准是 CART 算法的核心,从众多分组变量中找到最佳分割点,本文通过 Gini 指标来衡量数据纯度.

2.2　Gini 系数

假设样本数据分为 K 类,其中样本点属于第 k 类的概率为 p_k,则概率分布的基尼指数定义为:

$$\mathrm{Gini}(p) = 1 - \sum_{k=1}^{K} p_k^2. \tag{1}$$

对于二分类问题,若样本点属于第 1 个类的概率是 p,则概率分布的基尼指数为

$$\mathrm{Gini}(p) = 2p(1-p).$$

对于给定的样本集合 D,其基尼指数为

$$\mathrm{Gini}(D) = 1 - \sum_{k=1}^{K} \left(\frac{|C_k|}{|D|}\right)^2, \tag{2}$$

这里,C_k 是 D 中属于第 k 类的样本子集,K 是类的个数.

如果样本集合 D 根据特征 A 是否取某一可能值 a 被分割为 D_1 和 D_2 两部分,即

$$D_1 = \{(x,y) \in D \mid A(x) = a\}, D_2 = 1 - D_1. \tag{3}$$

则在特征 A 的条件下,集合 D 的基尼指数定义为

$$\mathrm{Gini}(D,A) = \frac{|D_1|}{|D|}\mathrm{Gini}(D_1) + \frac{|D_2|}{|D|}\mathrm{Gini}(D_2). \tag{4}$$

基尼系数 $\mathrm{Gini}(D)$ 表示集合 D 的不确定性,基尼指数 $\mathrm{Gini}(D,A)$ 表示经 $A = a$ 分割后集合 D 的不确定性,基尼指数越大,样本集合的不确定性也就越大.

2.3　决策树建立

输入: 训练数据集 D,停止计算的条件

输出: CART 决策树

根据训练数据集,从根结点开始,递归地对每个结点进行以下操作,构建二叉决策树:

(1) 设结点的训练数据集为 D,计算现有特征对该数据集的基尼指数. 此时,对每一特征 A,对其可能取的每个值 a,根据样本点对 $A = a$ 的测试为"是"或"否"分割为 D_1 和 D_2 两部分,利用式(4) 计算 $A = a$ 时的基尼指数.

(2) 在所有可能的特征 A 以及他们所有可能的切分点 a 中,选择基尼系数最小的特征及其对应的切分点作为最优特征与最优切分点. 依最优特征与最优切分点,从现结点生成两个子结点,将训练数据集依特征分配到两个子结点中去.

(3) 对两个子结点递归地调用(1) 和(2),直至满足停止条件

(4) 生成 CART 决策树

2.4　决策树剪枝

算法过程如下:

输入: CART 算法生成的决策树 T_0

输出: 最优决策树 T_α

1) 初始化 $\alpha_{\min} = \infty$,最优子树集合 $\omega = \{T\}$;

2) 从叶子节点开始自下而上计算各内部节点 t 的训练误差损失函数 $C_\alpha(T_t)$(基尼系数),叶子节点数 $|T_t|$,以及正则化阈值 $\alpha = \min\left\{\frac{C(T) - C(T_t)}{|T_t| - 1}, \alpha_{\min}\right\}$ 更新 $\alpha_{\min} = \alpha$;

3) 得到所有节点的 α 值的集合 M;

4) 从 M 中选择最大的值 α_k,自上而下的访问子树 t 的内部节点,如果 $\frac{C(T) - C(T_t)}{|T_t| - 1} \leqslant \alpha_k$ 时,进行剪枝. 并决定叶节点 t 的值,即概率最高的类别,得到 α_k 对应的最优子树 T_k;

5) 最优子树集合 $\omega = \omega \cup T_k\omega$, $M = M - \{\alpha_k\}$;

6) 如果 M 不为空,则回到步骤 4. 否则就已经得到了所有的可选最优子树集合 ω;

7) 采用交叉验证在 ω 中选择最优子树 T_α.

3　实验过程

3.1　数据处理及分析

(1) 数据定义

数据表中共包含十个字段,如表 1 所示,表中部分数据如表 2 所示:

(2) 缺失值处理及变量筛选

处理过程中未发现缺失值且未出现重复 ID,因此可直接进行变量筛选.

分析发现已流失客户的使用月数均小于 25,因此判定该字段并不关键,选择以下字段作为特征变量 extra_time,extra_flow,pack_type, pack_change, asso_pur、contract 以及 group_use,并将连续变量转换为二分类变量,将没有超出套餐的通话时间和流量记为 0,超出的记为 1. 转换后部分数据如表 3 所示:

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 1 相关字段及定义

字段名	定义
user_id	用户 id(用户的唯一标识)
pack_type	套餐金额(分三档,1 为 96 元以下,2 为 96 到 225 元,3 为 225 元以上)
extra_time	套外通话时长(单位:分钟)
extra_flow	套外流量(单位:M)
pack_change	是否曾经改变过套餐金额(0/1 ——1 代表是,0 则代表否)
contract	用户是否与运营商签订过服务合约(0/1 ——1 代表是,0 则代表否)
associated_purchase	用户在使用移动服务过程中是否还同时办理其他业务(1 = 同时办理一项其他业务,2 = 同时办理两项其他业务,0 = 没有办理其他业务)
group_user	是否为集团用户(0/1 ——1 代表是,0 则代表否)
use_month	使用月数
loss	流失情况(0/1 ——1 代表已流失,0 则代表未流失)

表 2 表中部分数据展示

user_id	pack_type	extra_time	extra_flow	pack_change	contract	associated_purchase	group_user	use_month	loss
1	1	792.83	-10.45	0	0	0	0	25	0
2	1	121.67	-21.14	0	0	0	0	25	0
3	1	-30	-25.66	0	0	0	0	2	1
4	1	241.5	-288.34	0	1	0	1	25	0
5	1	1 629.67	-23.66	0	0	0	1	25	0
6	1	182	-115.86	0	1	0	1	25	0
7	1	196.33	221.29	0	1	0	1	23	0

表 3 表中部分数据展示(处理后)

ID	套餐金额等级	是否产生额外通话时长	是否产生额外流量	是否有改变行为	是否有服务合约	是否关联购买其他产品	是否为集团用户	是否为流失用户
1	1	1	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0
3	1	0	0	0	0	0	0	1
4	1	1	0	0	1	0	1	0
5	1	1	0	0	0	0	1	0
6	1	1	0	0	1	0	1	0
7	1	1	1	0	1	0	1	0

(3) 相关性分析

通过相关性矩阵热力图观察各变量之间的相关性,可发现各属性间相关性较低.

3.2 建模过程

(1) 建立自变量 x,因变量 y 的二维数组;

(2) 以 7∶3的比例拆分训练集及测试集;

(3) 使用 CART 算法建立决策树模型并拟合训练,基于 Gini 系数进行分类,设置树的最大深度为 6,区分一个内部节点需要的最少的样本数为 9,一个叶节点所需要的最小样本数为 5;

(4) 模型评分值对比

针对测试集和训练集分别进行评分,如表 4,得知测试集和训练集评分值较接近,模型效果较好.

表 4 评分值对比

数据集	评分值
训练集	0.875 3
测试集	0.86

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

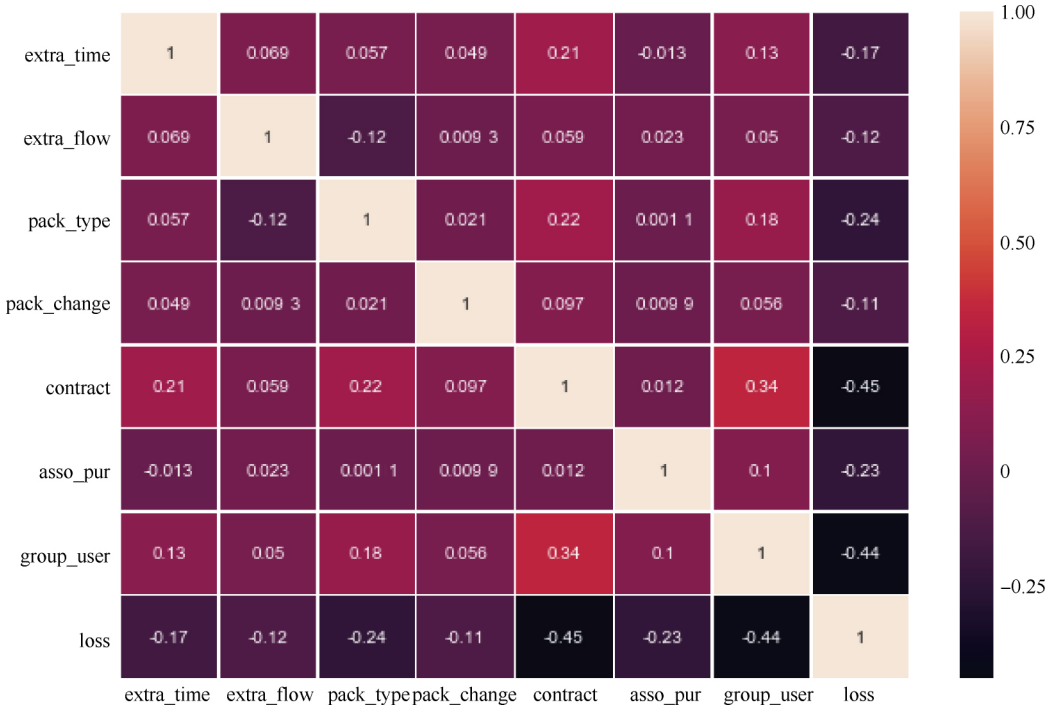


图 1　相关性矩阵热力图

3.3　模型优化

对于决策树来说,可调参数有

max_depth: 限定了决策树的最大深度,可有效防止过拟合;

min_samples_leaf: 限定了叶子节点包含的最小样本数,该属性可有效防止数据碎片问题;

min_samples_split: 分裂所需最小样本数;

min_impurity_split: 该值限制了决策树的增长,若某节点的不纯度(基尼系数,信息增益,均方差,绝对差)小于这个阈值,则该节点不再生成子节点.

由于本文训练数据有限,易出现过拟合现象,故期望能通过调整参数来有效防止过拟合. 本文选择对 max_depth 进行调整,缓慢增加深度来对模型进行训练,并计算评分数据,利用交叉验证法找出评分最高的索引. 同时输出模型参数与评分效果的关系图及最优参数值,如图 2 所示,优化后评分值对比如表 5 所示.

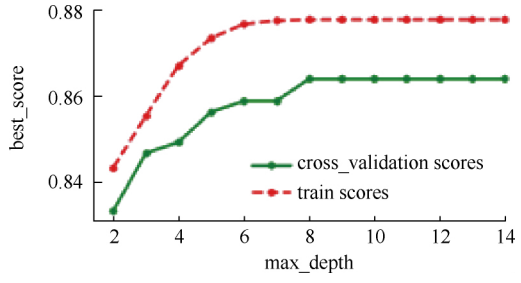


图 2　模型参数与评分效果的关系

表 5　评分值对比(调参后)

数据集	评分值
训练集	0. 865 3
测试集	0. 86

3.4　模型评估及可视化结果

输出决策树模型评价结果,如表 6 所示. 可得出建立的预测模型的精确率为 0. 86,说明在预测为流失的用户中,实际流失的用户占 86%; 同时召回率也为 0. 86,说明实际为流失的用户中,预测为流失的占 86%, $F1$ 值为 0. 86. 使用逻辑回归算法预测模型精确率为 0. 83,预测精度较低,且逻辑回归算法易欠拟合,对比可得出决策树模型算法的综合效果较优. 输出决策树模型的 roc 曲线图,如图 3 所示,可得 auc 值为 0. 73,模型还有待优化. 同时输出决策树结构图,如图 4 所示,部分决策树详情可见图 5.

表 6　模型评价结果对比

	precision	recall	f1-score	support
0	0. 72	0. 62	0. 66	324
1	0. 90	0. 93	0. 91	1 169
avg/total	0. 86	0. 86	0. 86	1 493

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

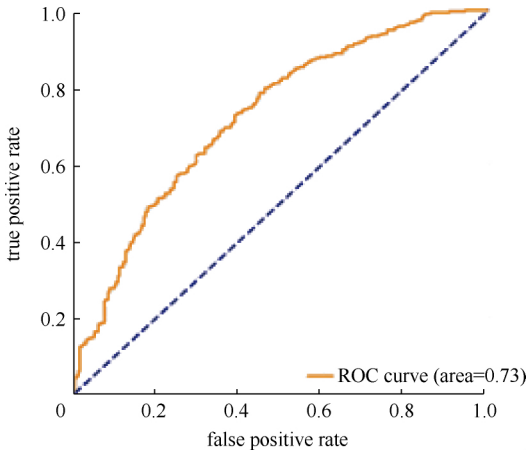


图 3 roc 曲线图

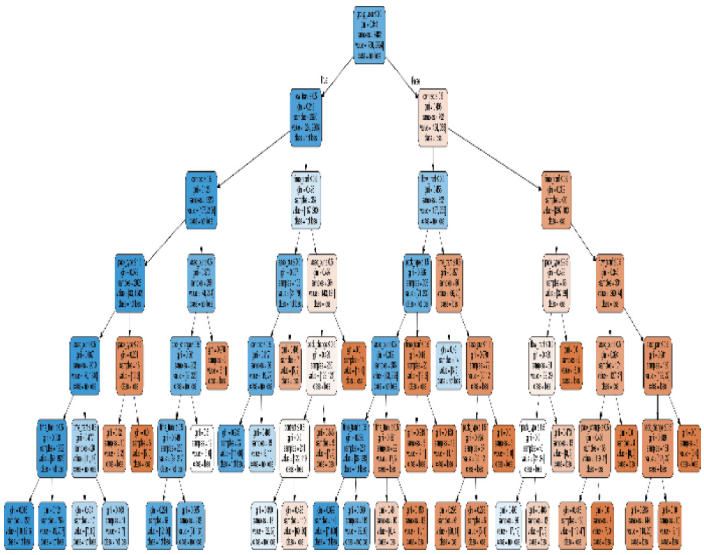


图 4 决策树结构图

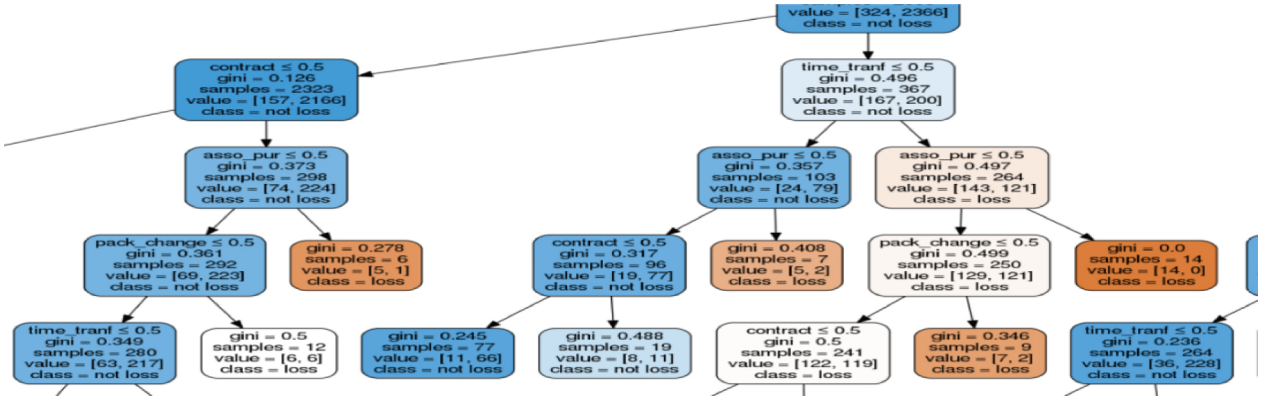


图 5 决策树部分分支展示

4 总 结

本文通过决策树中的 CART 算法成功预测到即将流失的客户,预测成功率达到 0.86,与逻辑回归算法相较更优,且 AUC 达 0.73,证明模型效果较好;但实际应用中要尽可能的找全部实际将流失的用户,即实际流失的用户中,模型能准确预测到的客户规模.

(下转第 34 页)

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[15] ZHAO J M,HUO S Y,HUANG H B,et al. Topological interface states of shear horizontal guided wave in one-dimensional phononic quasicrystal slabs [J]. Physica Status Solidi-Rapid Research Letters,2018,12(10):1800322.
[16] CARMICINO C, DI MARTINO G D. Analytical modeling of the one-dimensional acoustic field in hybrid rocket combustion chamber [J]. Journal of Sound and Vibration,2018,437:180 - 197.

Interface State in One Dimensional Photonic Crystal Heterostructures

GUO Yihang ZHAO Liming

(Department of Physics,Capital Normal University,Beijing 100048)

Abstract: The interface state of one-dimensional photonic crystal heterostructures $(AB)_m(CD)_m$ is studied by Dirac point method. The limit conditions of the existence of the interface state are discussed by adjusting the parameters of the system. The range of existing interface states is studied, and the relationship between the transmission of heterostructures states with refractive index under symmetric conditions is given. It is found that the corresponding optimal $(CD)_m$ structure can be found for each set of fixed $(AB)_m$ and the corresponding transmission can reach 1.

Keywords: one dimensional photonic crystal; heterostructures; topological interface state

---

(上接第 18 页)

在今后预测模型的优化上,还有很多的改进之处,如调整决策树的参数,特征的精细化筛选,或采用随机森林、遗传算法多算法相结合的方式进行模型评估,并分析用户分类、生命周期以及各变量之间的交叉性和相关性等等,以达到更好的预测效果.

参 考 文 献

[1] 李爱民. 基于逻辑回归和聚类分析的广电客户流失预警 [J]. 信息与电脑(理论版),2016(9):29 - 30 + 35.
[2] 周静,周小宇,王汉生. 自我网络特征对电信客户流失的影响 [J]. 管理科学,2017,30(5):28 - 37.
[3] 陈纪铭. 基于朴素贝叶斯分类算法的学员流失预警应用研究 [J]. 福建电脑,2017,33(8):105 - 106.
[4] 林明辉. 基于 BP 网络的通讯行业客户流失预警模型研究 [J]. 陕西学前师范学院学报,2016,32(3):146 - 149.
[5] HUANG J F, HE L L. Application of improved PSO-BP neural network in customer churn warning [J]. Procedia Computer Science,2018,131:1238 - 1246.
[6] 唐立,李六杏. 基于改进 CART 算法的 M-learning 过程中知识掌握程度预测 [J]. 韶关学院学报,2018(9):26 - 31.

Early Warning Model of Stock Customer Churn Based on Random Forest

YANG Rong ZHAO Juanjuan JIA Guojun

(Shanxi Normal University,School of Mathematics and Computer Science and Technology,Computer Science and Technology, Linfen Shanxi 044600)

Abstract: This study established a decision tree model based on the user package tariffs, and used the cross-validation method to select the optimal decision tree, further explored the lost user attributes, and provided decision support for operators to accurately initiate early warning retention and maintenance strategies, and got better experimental results.

Keywords: stock customer; decision tree; CART algorithm; early warning retentio

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net