基于决策树自动化特征选择的基金客户流失预测研究
——后疫情时代下的思考

董纪阳

(东北财经大学 管理科学与工程学院,辽宁 大连 116023)

[摘要] 客户是基金行业争夺竞争优势的焦点,对基金流失客户的准确预测是客户维挽和运营的前提,能大大增加自身的竞争优势。特别是 2020 年初暴发的新冠病毒肺炎疫情,给金融业发展带来了新的困难和挑战,当疫情跨地域跨国界蔓延时,其所致危机的复杂性更增加了经济前景的不确定性。流失客户往往具备一定的特征,机器学习方法被广泛应用于识别并预测即将流失的客户,但在用户特征的选择上往往过于主观。本文提出一种基金交易场景下,使用决策树方法对客户交易行为进行特征提取的方法,能够有效避免主观化特征筛选,大幅提升召回率。

[关键词] 决策树; 特征选择; 基金客户; 客户流失; 监督学习

[中图分类号] C934 [文献标识码] A [文章编号] 1003-4145 [2020] 09-0074-07

DOI:10.14112/j.cnki.37-1053/c.2020.09.012

一、问题的提出:后疫情时代下的基金交易与客户流失

客户是企业盈利的源泉,客户的忠诚度与客户关系的维持是各企业争夺竞争优势的焦点。自 20 世纪 60 年代,以客户为中心便成为企业经营的主旨。进入 21 世纪,全球历经了两次大的世界性金融危机以及数次公共卫生事件的冲击,其"蝴蝶效应"仍在加剧,人类已经步入"新全球危机时代"。特别是 2020 年初暴发的新冠病毒肺炎疫情,给金融业发展带来了新的困难和挑战,当疫情跨地域跨国界蔓延时,其所致危机的复杂性更增加了经济前景的不确定性。后疫情时代,以数字营销为代表的数字经济将成为发展主流,利用 AI 技术提供精准客户数据分析,时刻抓住客户的动态,将为企业行为模式上争取领先地位。客户流失管理是客户关系管理的重要环节,如何预判哪些客户有流失倾向,分析他们的流失原因,及时采取措施加以挽留意义重大。

客户流失预测一般指有针对性地对与当前公司表现出结束商业关系倾向的客户进行计算机化搜索和识别。近年来蓬勃发展的计算硬件与机器学习算法推动了客户流失预测算法在电信、银行、保险等应用场景的广泛应用,模型取得的斐然效果给企业带来了巨大收益,客户流失预测成为机器学习的重要应用方向。客户流失预测有着重大商用前景和学术价值,很多学者以客户或交易记录为数据集,运用多种机器学习方法展开了系列研究:Ganesh J 等用 SMOTE 算法进行数据均衡,选用决策树模型对信用卡数据进行客户流失挖掘①;Hung 等使用反向传播算法对台湾某通信公司的客户数据进行流失预测,论证了在各细分市场上建模的效果比在全部市场上更为准确②;还有学者使用基于粒子群优化(PSO)的欠采样与降维技术处理不平衡数据,构建随机森林模型对通信行业的流失预测模型。③

收稿日期:2020-06-01

作者简介:董纪阳(1979—),男,东北财经大学管理科学与工程学院讲师,管理学博士,主要研究方向为微观经济管理与决策。

①Ganesh J,Arnold M J,Reynolds,K.E.Understanding the Customer Base of Service Providers: An Examination of the Differences Between Switchers and Stayers,in *Journal of Marketing*, 2000,pp.65-87.

②Ha H,The moderating roles of status of B2B evaluator and dependence in the switching costs-switching intentions-performance causal chain in *Korea*,*Asia Pacific Business Review*,2017,pp.420-437.

③Kumar V,Reinartz W,Creating Enduring Customer Value,in *Journal of Marketing*,2016,PP36-68.

流失预测问题大都可以转化成二分类问题，分类预测模型本身的思想和应用在后续的建模与评估上与“客户预测”这个应用场景关联并不大。基于分类系统的客户流失预测模型的效率依赖于对可用数据集的学习策略。适当的预处理数据集有助于分类器达到所需的精度，最终获得理想的性能①。基金公司通过获取大量客户信息来归档数据，遗憾的是此类数据具有维度高与特征分布严重失衡的特点，流失客户数量通常与非流失客户相比要少得多，从而导致数据集不平衡。高质量的数据预处理对模型效果影响重大，在流失预测场景中常用的方法有数据均衡、人工特征选择、缺失值处理、特征降维等。人工特征选择的方法有较大主观性，本文在前人研究的基础上，使用决策树的方法进行特性选择，运用逻辑回归算法对流失预测的效果进行实验和对比评估，以期为后续相关研究提供参考。

二、客户流失预测模型构建的流程

对于一般机器学习，分类预测的流程通常需要数据录入、数据清洗、特征提取、特征筛选、模型训练、模型评估等步骤。数据清洗主要是将从信息系统中导出的数据进行一定程度处理，去除不标准数据和一些无用、杂乱的数据。特征提取、特征筛选用于提取能够体现数据特点的特征，模型训练主要是将数据提供给模型算法，让模型能够学习到一组参数，模型评估用来对模型的准确程度给予评价，看模型是否达到了一定的指标。

1.数据获取与数据清洗

在面向某个领域的数据分析任务时，首先需要确定能够获取的数据和数据的格式，这是数据分析的起点，之后针对每条数据来确定是否有确定的标签，如果有，就属于有监督学习；如果没有，则属于非监督学习。再进而确定是监督、非监督学习中的哪个具体的算法，或者归为某种具体的经典数学问题。②

在数据获取上要充分考虑数据的量级，如果数据量过大，可以考虑采用抽样来缩减处理的数据量，用样本代替整体；考虑后续特征抽取的数量，也可以对相关的特征进行压缩、降维；或者直接采用分布式引擎。在样本的获取上要做到多标签样本均衡，这尤其会影响分类问题的准确度。本文中面向的场景中，流失用户比例较低，那么就要对这类数据进行丰富，采用相关的数据平衡方式——过采样或者欠采样来应对。

特征工程的范畴较广，也是数据处理中较为耗时的一个阶段，是机器学习中基础而又必备的步骤，其中包括特征提取、特征选择、特征构建等方面。特征工程能够从繁杂的数据表现中，提取出面向问题最具代表性的那些特征，好的特征工程结果往往能够让简单的模型有较高的准确度，甚至高于复杂模型。对于分类模型，训练集中可能会存在某个或某些类别下的样本数远大于另一些类别下的样本数目，一方面需要对训练集做数据均衡，以提升模型对少数类的识别精度，另一方面不能简单地使用 AUC 指标衡量模型性能，而需要结合精准率召回率等多种指标。

常用的数据均衡方法有增加数据集、对多数类样本欠采样与对少数类样本的过采样。直接增加数据集效果显著，然而往往难以实现。欠采样是对多数类的数据样本进行采样来减少该类数据样本的个数，最直接的方法是随机地去掉一些多数类样本来减小多数类的规模，但是会丢失多数类样本中的一些重要信息，且不适用于少数类过少的情况。过采样指对少数类的数据样本进行采样以增加少数类的数据样本个数，最直接的方法是简单复制少数类样本或者使用 SMOTE 算法增加样本个数。

机器学习模型训练是一个持续优化的过程，需要经历模型训练、评估、参数调优的过程③。在训练过程中，通过绘制 loss 值曲线，能够判断模型是否已经收敛，为了避免过拟合，应该采用 K 折交叉验证，即将训练集分割为 K 个等分，每次训练从中选取一份作为测试机，其他作为训练集，这样对模型训练 K－1 次之后，取七个模型平均的 loss 值作为评估标准，就能够避免有偏采样作为测试集导致的欠拟合、过拟合问题，通过增加惩罚项、增加训练数据量等，也可以降低模型复杂度。

在训练后需要对模型的准确度进行评估，对于连续值可以采用距离计算，而布尔值可以采用混合矩阵方式来评估模型的准确性，业界通常采用 f 值计算来对一个模型的查准率、查全率进行评估。评估模型后，再次对参数进行调整，并观察 loss 值变化，直到可以收到满意的模型准确度。这是一个反复迭代的过程，可以通过人工经验来调整参数，也可以通过自动化方式对参数进行矩阵搜索尝试。在静态数据调优之后，将会把

①张线媚《数据挖掘在电信行业客户流失预测中的应用》，《微型机与应用》2015 年第 15 期。

②卢美琴、吴传威《大数据背景下商业银行贵宾客户流失的组合预测研究》，《电子商务》2019 年第 6 期。

③雷海锐、高秀峰、刘辉《基于机器学习的混合式特征选择算法》，《电子测量技术》2018 年第 16 期。

模型部署到线上使用，实际应用场景中数据条目、数据量也是动态变化。因此模型需要持续不断学习已有的数据，更新参数。

2.预测模型的特征提取

在机器学习中，特征是指实体的一些属性和性质，无论这些属性和性质是否对解决问题有用。在现实问题中，数据中的特征对于模型的训练和预测非常重要，更好的特征能够让模型简单而灵活。因此如何选择较好的特征是机器学习中重要的一环。特征选择分为特征提取和特征筛选两部分。在面向用户流失场景的分析时，用户本身的特性如性别、平均收入、年龄为静态数据，可以通过数值化、离散化的方式进行提取，特征提取后通过一定度量方法筛选出少量且能够保留大部分信息的特征，一方面可以减少特征数量、达到降维效果的同时使模型拥有更好的通用性和鲁棒性，减少过拟合；另一方面可以增强对特征和特征值之间的理解，提升模型的准确度。

从特征是否差异显著、特征与目标的相关性两个角度审视特征的价值是进行特征选择分析的有效途径。如果一个差异不显著，即该方差接近于 0，可以认为该特征不能对样本进行有效区分，没有太多的信息量；而对于与目标相关性不高的特征也应考虑删除，减少对模型的干扰。特征提取的手段很多，从面向特征的差异、相关性分析性角度出发，通常可以归纳为 Filter、Wrapper、Embedded 三类方法。

Filter 方法没有使用结果错误率来对特征选择的优劣进行打分，而是使用一些代理指标。这些代理指标通常计算速度更快，常用的代理指标包括逐点互信息、互信息、皮尔森积距相关系数。Filter 方法特征选取计算量一般比 Wrapper 要小。因为排除了使用模型对结果预测并评估准确度的步骤，选取的特征和使用某个模型无关，这也就使得特征更加通用，也更侧重特征之间的相互关系，但负面效果是会降低实际预测结果的准确度。此种类型的特征选择方法所产生的结果是对所有特征的打分排名，而不是某一组特定的特征组合，通过交叉验证，能够最终确定打分的截断值。在面向大量特征的时候，Filter 方法作为 Wrapper 的前置方法对特征进行预筛选，计算速度快使得它能够快速减少特征的数量。① Wrapper 方法使用预测模型来为特征选择子集打分。每次选择一组特征子集训练模型，之后对模型的预测结果进行打分，由于需要对特征的每种组合都训练一个模型，这会使得计算量非常大，但因为覆盖足够全面，较为容易找到合适的特征组合。Embedded 方法同样使用预测模型构建来选择特征，与 Wrapper 过程不同的是，在每次构建模型之后，对特征的权重进行分析。通常在模型构建时，加入惩罚项，L1 惩罚项会让某些低权重特征的权重倾向为 0，权重非 0 的特征会被选中；也可以用树模型，越靠近根、分叉越早的特征代表性也越强。

3.预测模型的特征筛选

从已经获得的特征中找出最有效的那一类特征就是特征筛选。一方面要能够代表实体的特性；另一方面，由于特征之间可能存在一定的关联关系，也需要对这些关系进行识别。本文采用计算协方差的方式：机器学习模型预训练，通过在已知数据上构建机器学习模型，一些模型可以获得每个特征所属的权重，通过按照特征对应权重由大到小排列，可以获得特征重要性排名。理论上通过碎石图可以帮助得到选择选取哪几个机器学习变量可以保留较多的信息量，在实际生产中，往往采用多次尝试构建机器学习模型的方法，不断减少特征来权衡精准率与特征数量之间的平衡。在本文中选择了决策树作为这种衡量特征重要程度的算法，决策树是一个有向无环图，树形结构代表实体属性和实体值之间的一种映射关系。树的每个节点标识一个对象，树杈代表了这个对象的取值范围的一次划分，叶子节点代表数据的一组分类结果。对应每条训练数据，都可以沿着根节点根据分叉条件逐层向下找到一条路径，到达最终的分类。建立树的过程是机器学习的训练流程。每个决策树都表述了一种树型结构，只是由它的分支来对此类型的对象依靠其属性进行一定的分类。每个决策树能够依靠对源数据的分割进行数据测试，这样能够使用满足划分准则的特征不间断地将数据集划分为信息纯度更高的子集。

其中不确定度的度量标准一般有信息增益、信息增益率、基尼指数三种。信息增益定义为熵与条件熵的差值，表征在某条件下信息不确定性减少的程度。对于待划分的数据集，其熵值固定，但是划分之后的熵就会有变化，熵越小表明使用此特征划分得到的子集的不确定性越小，因此两者的差异也就是信息增益越大，说明以当前特征划分后，信息纯度更高；如果某个属性存在大量的不同值，决策树在选择属性时会偏向于选

①杨荣、赵娟娟、贾郭军《基于决策树的存量客户流失预警模型》，《首都师范大学学报(自然科学版)》2019 年第 5 期。

择该属性，必然会带来较大偏差，信息增益率考虑了各分支数量的因素，定义为信息增益与数据集关于某特征的值得熵之比，其本质是在信息增益的基础之上增加了一个惩罚参数。特征个数较多时，惩罚参数较小；特征个数较少时，惩罚参数较大；基尼指数也叫基尼不纯度，表示在样本集合中一个随机选中的样本被分错的概率。集合所包含的纯度越高，集合里被选中的样本被分错的概率如果越小，它的基尼指数也就越小。①

决策树不仅可以用于模型构建，还可以用于特征筛选。决策树每次分叉都会选择对信息熵影响大的特征，所以我们将特征根据分叉的先后顺序排序，排序约靠前的特征就是对分类结果影响最重要的，通过这种方法筛选特征能够有效降低模型的复杂度。

4.监督学习的过程

在监督学习中，每条数据对的输入特征通常是一个向量，而确定的标签是一个值。模型训练后得到了映射函数，当把新的输入交给函数时，就会得到对新数据的一个预测结果。此时如果标签是一个连续值，就叫做回归问题，如果标签是一个枚举值，就叫做分类问题。通过对已有数据的观察，然后将此规律应用到新的数据上去，需要总结对问题足够通用的学习规律，这叫做模型的泛化能力。也并非漫无目的地去找寻这个映射函数，可以对问题给予一个基本的假定，然后推导出一个通用的公式，再通过现有数据来确定其中的参数。不同的假定也就产生了不同的模型，比如逻辑回归、支持向量机等。

下面将整个过程数学化表示，给定的数据为 $(x, g(x))$，其中就是目标函数。假设符合 g 行为的样本是从某个空间中，以未知概率 p，以独立同分布随机方式来抽样。这时定义一个损失函数。

$$L: Y \times X \longrightarrow R$$

其中，Y 是 g 的陪域，如果 g 预测出的值是 z，观测真值是 y，定义 $L(z, y)$ 叫为损失值，L 取值一般为非负实数。假定 p 是离散的，在全部样本上的损失值累计为：

$$R(f) = \sum_{i=1}^{n} L(f(x_i), g(x_i)) p(x_i)$$

那么问题简化为，如何确定函数 f^*，能够使得 $R(f^*)$ 风险值最小。根据 g 可以适用于全部观测值对 (x_1, y_1)，…，(x_n, y_n)，则以一种近似方式给出风险值的计算方式如下：

$$\widehat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} L(f(x_i), y_i) \quad (1)$$

通过统计理论就可以验证经验风险最小化是否可行，以及确定最小化的值。这就可以确定 f^*，从而找到 $(x, g(x))$ 的一种风险最小化的映射关系。

三、基金公司客户流失模型构建

我国的基金市场发展近三十年，其技术环境、监管环境、政策环境得到不断发展和提升，而基金公司内部的治理结构、管理人监督也随之完善，共同推进了中国金融体系的成长。在不断健全发展的大环境下，客户开始认可重视基金这种投资方式，影响客户基金投资决策的影响因素很多，一方面是经济形势、企业发展、行业发展、科学技术演进等宏观因素；另一方面是客户自身的投资条件、心理预期、风险承受、投资动机等微观因素。诸多因素影响着客户对基金的认识以及选择，尤其是后疫情时代的经济环境存在着诸多不确定性，而个人信息与交易信息能够在一定程度上反映投资特点，可以以此数据进行流失客户的识别。

（一）数据选取

本文以深圳市某基金公司的客户为研究对象。采用客户信息表和交易记录表作为数据来源，其中客户信息表是客户开户时填写的情况，交易记录表则是按时间顺序客户的一笔笔交易行为，这样的交易行为带有时间属性。受外部环境和自身投资习惯的影响，用户对基金的买进与卖出具有很大的变动性，客户流失的有效预判价值巨大。数据集为该公司 2018 年 5 月 1 日至 2018 年 12 月 31 日这 8 个月的交易信息表以及客户信息表的数据，如表 1。值得说明的是，本文所选用的数据虽然为 2018 年所收集的，但是对于后疫情时代下的金融发展有较好的启示和借鉴意义。

在经过特征提取后，得到了如下特征，令特征为 f_n，$n = 1, 2, 3\cdots$，对应上表中的特征得到：

$$f_1, f_2, f_3, \cdots f_{13}$$

①马文斌、夏国恩《基于深度神经网络的客户流失预测模型》，《计算机技术与发展》2019 年第 9 期。

表 1 客户交易数据表

序号	维度	特征	类型	备注
1	客户	编号(唯一标识)	数值	客户的 ID
2	客户	客户类型	类别	0:机构,1:个人
3	客户	邮编	数值	客户居住地邮编
4	客户	家庭地址	字符	客户居住地址
5	客户	性别	类别	0:女,1:男
6	客户	出生日期	时间	出生年、月、日
7	交易	交易日期	时间	交易日期和时间
8	交易	交易类型	类别	
9	交易	基金代码	类别	
10	交易	渠道	类别	
11	交易	网点	类别	指定网点
12	交易	该笔交易的购买份额	数值	买入为正,卖出为负,0 可忽略
13	交易	剩余份额	数值	

客户的流失按照是否已经完全流失,一般可分为已经流失与正在流失两种。对前者可以直接通过统计方法找出持仓量一直为 0 的流失用户 $f_{13}=0$,无需识别预测;后者是模型预测关注的重点,可以通过统计某段时间的增仓、减仓次数,定义增仓数目为 0,减仓数大于 0 的用户为流失用户。

本文使用前 6 个月的上述特征数据作为模型输入 $f_1,f_2,f_3,\cdots f_{13}$,后 2 个月数据提取出流失标签 $target\in(0,1)$,流失定义为 1,未流失为 0,构建监督学习模型对基金客户流失进行预测,目标就是找到合适的 F。

$$F(f_1,f_2,f_3,\cdots f_{13})\rightarrow target$$

在对数据进行缺失值填充和正负样本平衡后,将数据进行 OneHot 编码:

$$f_1\rightarrow f_{11},f_{12},f_{13}\cdots f_{1n}$$

上述机器学习问题变换为:

$$F(f_{11},f_{12}\cdots f_{1n},f_{21},f_{22}\cdots f_{2m},f_{31},f_{32}\cdots,f_{3q})\rightarrow target$$

其中 n,m,q 代表 OneHot 编码之后的特征数量,在变换后得到 914 个特征。

以上问题等价于:

$$F(\dot{f}_1,\dot{f}_2\cdots,\dot{f}_p)\rightarrow target$$

其中 p=914,等价于 914 个特征。

随着特征迅速膨胀,将这些特征全部放入模型训练过程,将会使得训练流程变得冗长且非常容易过拟合。

(二)决策树特征筛选

预处理后最终得到基金客户流失预测的数据,训练集 $train\_total\_sample$ 3086 条,其中正负样本数各 1543 条,测试集 2693 条,基本满足实验需求。对于 F 这里使用决策树来进行模型构建,考虑到基尼系数在大幅减少对数运算的基础上保持熵模型的优点,本模型的度量标准选择基尼系数。由于特征数量很多,选择决策树中的给与枝剪策略,树深度控制在 200。

$F(Gini,Deepth<200)$

对于决策树,每个节点由多个属性组成,见图 1:

叶子节点 $Leaf$:

分叉属性 $feature$,对应本实验中的 $\dot{f}_1,\dot{f}_2\cdots,\dot{f}_p$

信息纯度 $gini$:根据决策树计算的信息纯度

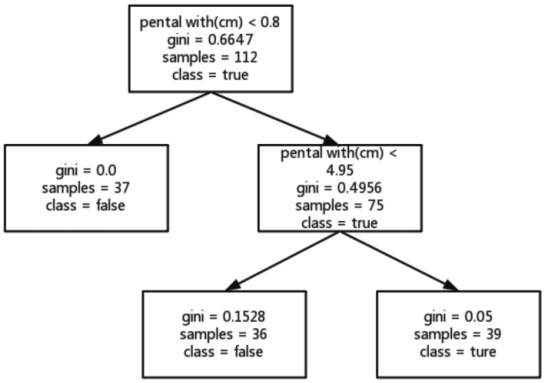

图 1 决策树节点属性

$$gini=\sum_{i=1}^{p}\dot{f}_i(1-\dot{f}_i)$$

此节点下的样本数 $sample$,本实验中

$$sample\in(0,3086)$$

此节点下对于属性的样本类别 $class$

对于正样本,即流失用户样本 $class=true$;对于负样本,即未流失用户样本 $class=false$

非叶子节点 *Non-Leaf*: 非叶子节点中没有分叉属性，其余和叶子节点一致。

因为分叉 *feature* 都在叶子节点上，

定义: *Leaf* 的左子树 *left*，*Leaf* 的右子树 *right*，对任意特征 $\hat{f}_i$，特征重要度 $feature\_importance_i$ 计算方式如下:

$$\begin{aligned} feature\_imp\ ortance_i &= (Leaf.sample^* \ Leaf.gini - left.sample^* \ left.gini \\ &- right.sample^* \ right.gini) / train\_total\_sample \end{aligned}$$

$feature\_importance_i \in (0,1)$

本实验中，保留 99%的特征信息，所以对 $feature\_importance_i < 0.01$ 时，认为该特征的信息不足以表达足够信息，去掉这些特征。

经过筛选，914 个特征保留了 113 个。追溯这些特征的含义，增仓减仓标签有着比较高的重要性，与定义的流失标签有较大关联，其他较高的特征为网点、基金代码、持有份额、(交易) 天、省份、城市、街区等特征。从数据上看，交易信息的重要性略大于个人信息，在特征筛选中占有更大的权重。网点体现了较强的地域特征，表明交易地点对流失有较大的影响; 基金代码与持有份额的变化是客户对基金预期直接体现。

(三) 流失模型构建

在流失预测模型的构建上，本文使用的算法有较有代表性的逻辑回归。逻辑回归是一种用于解决二分类问题的机器学习方法，用于估计某种事物的可能性。其数学模型、求解和实现都相对简洁。逻辑回归以线性回归为理论支持，通过引入 Sigmoid 函数将预测值映射在将数值结果转化为了 0 到 1 之间的概率，从而通过阈值实现二分类。考虑到数据已进行过特征选择，惩罚系数选择 L2 正则化，选择 liblinear 优化算法，通过坐标轴下降法来迭代优化损失函数。

在经过特征筛选之后问题简化为:

$$F(\hat{f}_1, \hat{f}_2 \cdots \hat{f}_{113}) \rightarrow target$$

根据逻辑回归模型基本假设

$$target = \theta_1 \hat{f}_1 + \theta_2 \hat{f}_2 + \cdots + \theta_n \hat{f}_n = \theta^T X, n = 113$$

$$令\ h_\theta = \frac{1}{1 + e^{-target}} = \frac{1}{1 + e^{-\theta^T X}}$$

h_θ 的实际意义为样本对应的 target 的二分类概率:

$$P(target = 1 \mid \hat{f}; \theta) = h_\theta(x)$$

$$P(target = 0 \mid \hat{f}; \theta) = 1 - h_\theta(x)$$

下面进行极大似然估计计算，概率函数为:

$$P(target \mid \hat{f}; \theta) = (h_\theta(x))^{target} * (1 - h_\theta(x))^{1-target}$$

因为样本数据独立，所以联合概率分布函数可以表示为各个边际分布的乘积，取似然函数为:

$$L(\theta) = \prod_{i=1}^{m} P(target^{(i)} \mid \hat{f}^{(i)}; \theta)$$

取对数似然函数:

$$l(\theta) = \log(L(\theta)) = \sum_{i=1}^{m} target^{(i)} \log(h_\theta(x^i)) + (1 - target)^{(i)} \log(1 - h_\theta(x^i))$$

$$J(\theta) = -\frac{1}{N} l(\theta)$$

最大似然估计为使得 $l(\theta)$ 取最大值时候 θ 的值，这里可以用梯度上升法来求解，取

$$\frac{\partial J}{\partial \theta} = -\frac{1}{n} \sum_{i=1}^{m} (target^{(i)} - target^{*(i)}) x^i + \lambda \theta$$

这样就得到了一组 θ 从而求得 F。

(四) 模型结果的效果评估

二分类模型的单个样本预测有四种结果，这四种结果可以写成一个 2* 2 的混淆矩阵，如表 2 所示，用 T

(True)代表正确、F(False)代表错误,TP 与 TN 表示预测值与实际值相符,模型预测正确。而 FP 与 FN 表示预测值与实际值不符,模型预测错误。

表 2 混淆矩阵

	预测为正类	预测为负类
实际为正类	TP	FN
实际为负类	FP	TN

以混淆矩阵作为基础,我们选择精准率、召回率和 F1 值作为分类模型的评价指标。其中,所有样本能够被正确预测的比例称为精准率(公式 2),实际为正类的样本中能够被正确预测为正类的比例称为召回率(公式 3),F1 值用精准率和召回率的调和平均数表示(公式 4)。考虑到基金客户流失的目的在于准确识别潜在流失客户,所以本文关注的重点在召回率和 F1 值。

$$P = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

$$R = \frac{TP}{TP + FN} \quad (3)$$

$$\mathrm{F1} = \frac{2TP}{2TP + FP + FN} \quad (4)$$

对训练的评估模型进行检验(见表 3),在本实验中分别对使用决策树筛选的特征建模 M1,和未经决策树筛选的特征采用逻辑回归建模 M2。可以看到 M1 分类器的效果较好,在对正类的识别上,预测为正类的全部是正类,在对负类的预测上相对较好,预测为负类的有 32 个实际是负类,有 62 个负类样本没有识别出,精准率和召回率分别为 0.99 与 0.98,F1 值为 0.98,都为较高水平。M2 分类器的效果不理想,在对负类样本的预测上,只识别出 25 个负类样本,而将 598 个正类样本预测为负类,精准率和召回率分别为 0.99 与 0.78,F1 值为 0.86,与 M1 分类器相比,精准率差距不大,召回率差异显著,处于较低水平,该分类器无法识别负类样本。

表 3 基金客户流失分类模型预测结果

方法	混淆矩阵		精准率	召回率	F1 值
(M1)逻辑回归决策树特征筛选	TP = 2599	FN = 62	0.99	0.98	0.98
	FP = 0	TN = 32			
(M2)逻辑回归全特征	TP = 2063	FN = 598	0.99	0.78	0.86
	FP = 7	TN = 25			

实验结果表明,未经过特征筛选的分类模型在对正负样本严重失衡的数据集中效果不佳,体现在无法有效识别负例样本,而负例样本恰恰是我们重点关注的。而经过决策树筛选后,都能够在测试集上较为精确区分正例(未流失客户)和负例(流失客户),最高能够达到了 99%的精准率和 98%的召回率,因此利用分类模型可以在流失进行有效的预测。

四、结论

准确的客户流失预测是客户维挽的前提和基础,本文提出一种基金交易场景下,使用决策树方法对流失客户特征自动化筛选的方法。以某基金公司的客户和交易两个维度的数据为例,进行特征提取和决策树特征筛选,发现交易信息对模型有着更高程度的影响。对流失影响较大的特征为网点(地域)、基金代码、剩余份额。分别使用经过决策树筛选的特征组和未经决策树方法筛选的特征组通过逻辑回归算法构建流失预测模型,对使用混淆矩阵、精准率、召回率、F1 值指标其上述模型的效果进行评估。本特征自动化提取方法可以较为准确的提取对目标信息贡献度较高的特征,大幅提升召回率。数据挖掘技术是客户流失精准预测的支持,而个性化的营销维挽是最终项目落地的关键,需将两者有机结合,实现更高水平的金融服务。总之,基于 AI 技术的客户流失预警将快速调整企业流程并保持客户满意度,从而提高了客户忠诚度和保留率,将成为以基金业为代表的金融行业应对后疫情时代条件下客户管理的对策和良方。

(责任编辑:陆影)