DOI:10.13878/j.cnki.jnuist.2016.03.009

来鹏[1]　赵茹蕾[1]　郭利珍[1]

银行客户定期存款认购的统计决策研究

摘要

当今银行之间的竞争日益加剧,能有效地挖掘潜在客户并为之提供差异化服务,对提高银行竞争力尤为重要.用决策树算法对可能影响银行客户是否认购定期存款的21个影响因素进行数据挖掘分析,构建了银行客户认购定期存款业务影响因素的决策树模型.研究结果表明显著影响客户认购定期存款的3个因素为员工指标人数、持续时间和月份,这可以大大缩小银行推送认购定期存款的客户范围,有利于提高银行效率.

关键词

数据挖掘;客户定位;决策树;统计决策

中图分类号 F830

文献标志码 A

0　引言

近年来,随着外资银行全面进军中国银行业并且逐步成为我国银行体系中的重要力量,我国银行业间的竞争空前激烈,而要在市场中立于不败之地,就要提高客户事务的处理能力,对客户进行深层次挖掘以及合理的定位,实现高效管理[1].因此,客户的合理分类是提高银行客户管理效率的基础和前提.

本文以葡萄牙银行机构提供的客户数据为研究对象,讨论影响银行客户认购定期存款的因素,对可能认购的用户进行客户定位,便于银行提高工作效率,更好地为客户提供服务.注意到所研究的问题从本质上是一个分类问题,是通过多个研究因素判定客户是否为具有效益的优质目标客户,而从数据类型来看,数据呈现出既有离散变量,又有连续变量,既有二值变量,又有多值变量等特点,很多传统的建模预测方法、分类方法不再适用.例如:线性回归模型由于其模型假设不再满足;非参数回归方法会面临维数过高的问题;神经网络模型又过于复杂,在计算效率上比较差;判别分析方法因为数据的复杂特性,很难确定合适的、符合复杂数据类型的恰当距离函数来构造判别准则.所以,决策树方法就由于其对数据类型的较弱假设,计算效率比较高,处理离散或复杂分类数据比较有效的特点而在本文中被采用.

1　基于信息熵的决策树算法

信息熵又称为期望信息量,是用来衡量信息量凌乱程度的指标,熵值越大,则代表信息的凌乱程度越高.基于信息熵的决策树算法是通过收集已知类别的样本,将提供最大信息增益的属性作为节点分裂方案去构造决策树的,即所选测试属性是从根到当前节点的路径上尚未被考虑的具有最高信息增益属性.决策树的每个节点对应一个非类别属性,每条边对应该属性的每个可能值[2].

设 S 是 s 个数据样本的集合,不妨设类标号属性具有 n 个不同的值,定义 n 个不同的类为 $C_i(i=1,2,\cdots,n)$,s_i 是类 C_i 中的样本数.设一个属性 D 有 m 个不同的取值 $\{a_1,\cdots,a_m\}$,使用属性 D 可将样本集合 S 划分为 m 个不同的集合 $\{S_1,S_2,\cdots,S_m\}$,其中 S_j 包含了集合 S 中属性 D 取值 a_j 时的数据样本.若属性 D 被标记为测试属性,即用于对当前的样本集进行划分,设 s_{ij} 为样本子集 S_j 中属于类别 C_i 的样本数,那么

收稿日期 2015-07-02

资助项目 江苏省高等学校大学生创新创业训练计划(201410300013Z);2016年度江苏高校"青蓝工程"培养对象

作者简介

来鹏,男,博士,副教授,研究方向为复杂数据统计分析以及数据挖掘.

laipengnuist@163.com

1　南京信息工程大学　数学与统计学院,南京,210044

根据属性 D 划分当前样本集所需要的信息熵的计算公式为

$$E(D)=\sum_{j=1}^{m}\frac{s_{1j}+s_{2j}+\cdots+s_{nj}}{s}I(s_{1j},s_{2j},\cdots,s_{nj}),\quad(1)$$

其中,$\frac{s_{1j}+s_{2j}+\cdots+s_{nj}}{s}$ 可以作为第 j 个子集 S_j 的权值,它是由该子集中所有属性取值为 a_j 的样本数之和除以集合 S 中的样本总数而得到的,$E(D)$ 的计算结果值越小,表明子集划分的纯度越高.此时,对于子集 S_j 的信息量的计算方法为

$$I(s_{1j},s_{2j},\cdots,s_{nj})=-\sum_{i=1}^{n}p_{ij}\log_2(p_{ij}),\quad(2)$$

其中,p_{ij} 表示样本子集 S_j 中任意一个数据样本属于类别 C_i 的概率.因此,利用属性 D 对当前分支节点进行相应的样本集划分所获得的信息增益为

$$G(D)=I(s_1,s_2,\cdots,s_n)-E(D),\quad(3)$$

其中 $I(s_1,s_2,\cdots,s_n)=-\sum_{i=1}^{n}p_i\log_2(p_i)$,$p_i=s_i/s$,$i=1,2,\cdots,n$.换言之,$G(D)$ 就是根据属性 D 的取值进行样本集划分所获得的信息熵的减少量,决策树归纳算法用于计算每个属性的信息增益,从中挑选出信息增益最大的属性作为给定集合 S 的测试属性,并由此产生相应的分支节点.所产生的节点被标记为相应的属性,并根据这一属性的不同取值分别生成相应的(决策树)分支,每个分支都代表一个被样本划分的样本子集[3].

2 银行客户认购定期存款建模方案研究

2.1 数据介绍以及变量描述

现代的银行的客户关系管理,需要面对海量的客户信息,这就需要银行对数据库中的原始客户数据进行深层次的挖掘,寻找目标客户.所用数据集是葡萄牙银行机构从 2008 年 5 月—2010 年 11 月所有话访活动市场调查结果的 41 188 个银行客户的相关数据[4],用来预测银行客户是否认购其定期存款并将其分类.将记录的 21 个属性变量假定为影响顾客是否认购存款业务的影响因素.这些统计变量可分为 4 类:客户情况、与银行关系、接触银行活动状况和经济社会环境状况.具体表现为

1)客户情况:年龄、工作状况、婚姻状况、受教育程度、房贷、个人贷款.

2)与银行关系:信用拖欠状况、账户余额、认购定期存款情况.

3)接触银行活动状况:被联系的方式、近月接触日期、近年接触月份、联系的持续时间、本次活动期间被联系次数、之前接触次数、之前的活动结果、距上次联系过去的天数.

4)经济社会环境状况:员工人数的季度指标、就业变化率、消费者信心指数、居民消费价格指数.

2.2 数据的处理与转换

从对该葡萄牙银行机构的 41 188 个银行客户的相关数据的初步研究发现,该数据集数据量比较大,如果将全部数据用于分析,会发现数据过多使得计算效率比较低.一个简单的解决办法是通过随机采样的方法随机抽取部分数据,使数据具有足够的代表性,能够快速准确地得到正确的分析结果.在此基础上,本例将数据分割成训练数据集(70%)和验证数据集(30%),这样在用训练数据集建立好模型后,利用验证数据集对模型进行修正预测,从而避免模型的过度拟合,提高模型的灵活性,最终提高模型的质量和预测效果.

由于对本例通过数据初步了解客户最终认购的比例占到 12.7%,还有 87.3%的银行客户并没有响应,两种数据之间相差过大,如果直接对该数据进行建模,将由于两者数据量差别太大,可能使得分析结果有偏差,给模型的建立以及预测能力带来较大的负面影响.为了更好地进行建模,对数据进行更准确的分析,本文在最终认购定期存款的客户随机抽取 2 060 个样本形成 SAS 数据集 YES,在不认购的客户中随机抽取 2 060 个样本形成 SAS 数据集 NO.合并 YES 与 NO 数据集,使之变成本文最终所用的测试集 NEW,使最终认购的比例与拒绝认购的比例大致相等,从而使得各类数据的特点能更好地体现出来.然而,考虑到这种认购比率与现实生活的实际比率并不相符,抽取数据的结果并不能代表真实情况,所以为了考虑到原始数据之间的相互比例关系,基于贝叶斯原理的先验概率将被作用于目标变量,帮助我们将原始数据的先验信息加以添加,避免数据抽样后导致的不足,从而使研究结果适合用于解决实际问题.

在古典拟合模型中,通常是以变量服从正态分布作为基本假设,在变量为正态分布条件下,模型的拟合效果往往也比较好,具有比较好的分析性质.另外,如果变量的类别过多,观测样本又仅仅集中在少数类别中,那么合理的类别合并有助于提高建模准确性和估计效率.因此,对于一些分布很分散的连续型变量数据,可以通过函数变换的方式对其进行转

换,使其分布更贴近正态假设;对于多重分类变量,也可以通过合并来进行数据整理.本例对此类数据进行了如下处理:

对 duration 进行分组转换,将联系的持续时间分为[小于等于 373.6 s]、[大于 373.6 s]2 组,分组后的持续时间分布如图 1、2 所示.

此外,依次对 campaign 进行分组转换,将本次活动期间被联系次数分为[1~6 次]和[大于 6 次]2 组;对 Pdays 进行分组转换,将距上次联系过去的天数分为[1~60 d]、[大于 60 d]的 2 组;对 cons_price_idx(消费者价格指数)进行分组,按[小于等于 92.843]、[大于 92.843 且小于等于 93.64]、[大于 93.64]分 3 组;对 cons_conf_idx(消费者信心指数)进行分组,按[小于等于 35]、[大于 35 且小于等于 43.29]、[大于 43.29]分 3 组;对 Euribor3M(拆借利率每日指标)进行分组,按[小于等于 4.5]、[大于 4.5]分 2 组;对 nr_employed(雇员人数)进行分组,按[小于等于 5 029 人]、[5 030~5 161 人]、[5 162 人以上]分 3 组(此处不附图赘述).

2.3 银行客户认购定期存款建模结果分析

数据挖掘的目的是从数据中挖掘客户价值,目的不仅是要以此为例揭示如何合理地进行客户定位,更是为了提高银行的利润,从而使银行在以后的经营活动中,能够更加注重数据挖掘方法,将此技术运用到生产实践中去,最终提高银行的竞争力[5].本文以葡萄牙银行机构从 2008 年 5 月—2010 年 11 月所有话访活动市场调查结果为基础数据,在以错判损失最小化选择最优模型的原则下,建立进行一次电话访问的成本为 20 美元,成功之后的收入为 80 美元的收益矩阵,结合贝叶斯先验信息,构造了基于

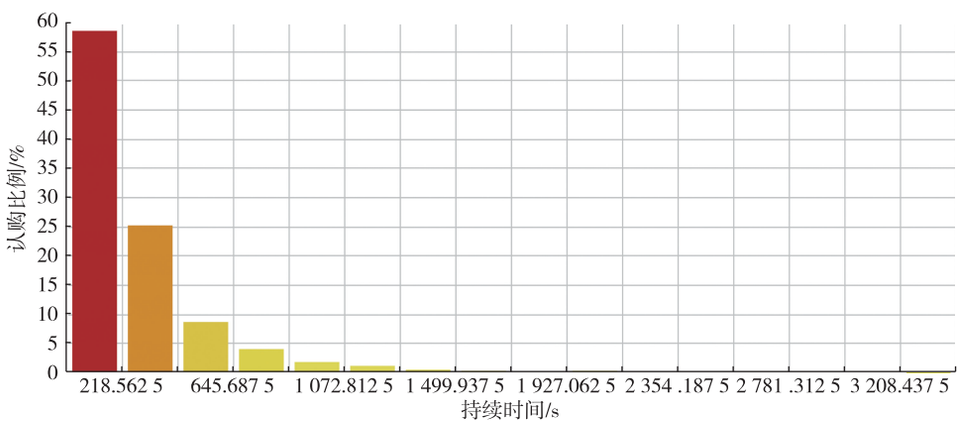

图 1 转换前 duration 的分布

Fig. 1 Distribution plot for variable duration before transformation

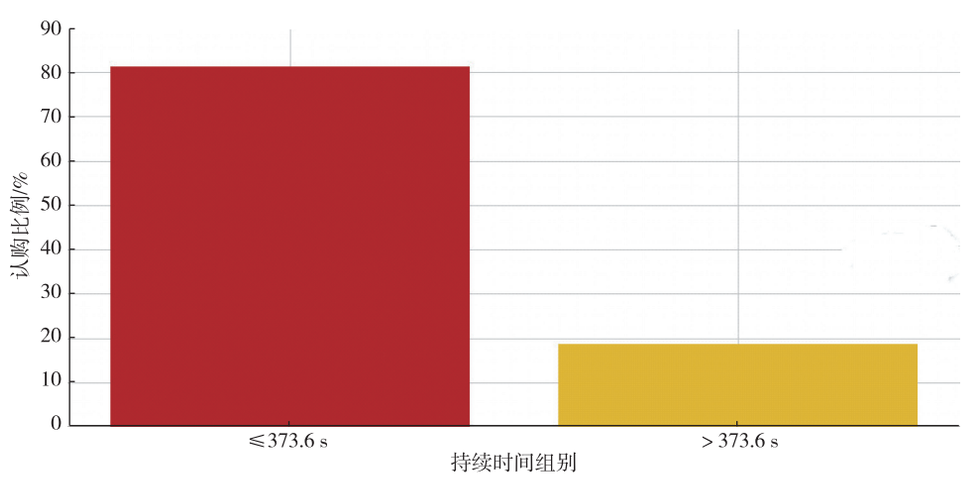

图 2 转换后 duration 的分布

Fig. 2 Distribution plot for variable duration after transformation

信息熵增益最大化的决策树收益最大化模型[6].通过对数据的分析处理,得到分析结果如表 1 和表 2 所示.

表 1　训练集(TRAIN)拟合结果

Table 1　Simulation results for training data

实际	预测	
	否	是
否	95%	22%
是	5%	78%

表 2　验证集(VALID)拟合结果

Table 2　Simulation results for validation data

实际	预测	
	否	是
否	92%	22%
是	8%	78%

在运用了先验概率 prior 的基础上,训练集(TRAIN)中预测结果为"是"实际也为"是"的概率达到了 78%,预测为"否"实际也为"否"的概率达到 95%;验证集(VALID)中预测结果为"是"实际结果也是"是"的概率达到了 78%,预测为"否"实际也是"否"的概率是 92%.训练集和验证集的结果几乎相同,可以看出,对客户是否认购定期存款得到了较好的预测.此外,为了根据数据所建立模型得到影响客户认购存款的关键因素,从决策树的分类结果中,还可以总结出下列重要规则(表 3),可以用于银行的实际目标客户定位.

规则 1—规则 2 表明在经济社会环境背景下,员工指标人数(nr_employed)相对越小客户认购定期存款的机率越大.在员工指标人数小于 5 161 时定期存款的认购比例达到 72.8%,而在员工指标数量大于 5 161 时定期存款的认购比例仅仅为 26.9%.由此可以发现经济社会背景中的员工指标人数因素对客户认购定期存款有影响作用.

从规则 3—规则 6 可以看到最近一次与银行接触持续时间的长短(duration)是影响客户是否有意认购定期存款的影响因素.联系持续时间长的客户认购定期存款的比例高于持续时间短的客户认购定期存款的比例.这表明持续时间同样是影响客户认购定期存款的重要因素之一.从规则 3—规则 4 看到在持续时间都大于 373.6 s 时员工人数的季度指标相对越小的客户认购定期存款的比例达到 90.3%,这进一步证明了员工人数的季度指标对定期存款认购的影响作用.

规则 7—规则 8 表明最近一次接触的月份(month)也是影响客户认购定期存款的一项因素.在月份划分中能够明显地看出 5 月客户认购定期存款的比例会降低,考虑可能与 5 月楼市回暖、各类投资理财迅速崛起以及银行一系列定向降准政策有关,这些因素会直接导致银行认购定期存款比例走低.

根据决策树预测模型的依赖关系可以发现,能够对预测属性产生影响的属性由强到弱依次是员工指标人数、持续时间和月份.因此本研究选取的 21 个可能影响定期存款认购因素中员工指标人数为最显著的影响,持续时间和月份次之,其他 18 个因素对客户是否认购定期存款影响并不显著.

3　结束语

将客户关系放到银行经营的核心位置,应当是银行的实际营销理念.而利用数据挖掘分析客户数据、掌握客户特征、挖掘客户价值,才能为企业带来显著利润[7].本文以葡萄牙银行机构从 2008 年 5 月—2010 年 11 月所有话访活动市场调查结果为基础数据,运用决策树信息熵的归纳算法进行数据挖

表 3　决策树模型的规则

Table 3　Rules for the decision tree model

规则序号	规则内容
1	在 nr_employed≤5 161 时,认购比例达到 72.8%.
2	在 nr_employed>5 161 时,客户有 26.9%的概率认购定期存款.
3	在 nr_employed≤5 161 时,duration>373.6 s 时,客户有 90.3%的概率认购定期存款.
4	在 nr_employed>5 161 时,duration>373.6 s 时,客户有 69.1%的概率认购定期存款.
5	在 nr_employed>5 161 时,duration≤373.6 s 时,客户只有 2.2%的概率认购定期存款.
6	在 nr_employed≤5 161 时,duration≤373.6 s 时,客户有 63.1%的概率认购定期存款.
7	在 nr_employed≤5 161 时,duration≤373.6 s,在 5 月以外的其他月份,客户有 76.5%的概率认购定期存款.
8	在 nr_employed≤5 161 时,duration≤373.6 s,在 5 月时,客户有 19.4%的概率认购定期存款.

掘,探究影响客户认购定期存款的影响因素,研究最终发现显著影响客户认购定期存款的因素只有员工指标人数、持续时间和月份3个指标,大大缩小了银行推送客户认购定期存款的客户范围,显著提高了银行的投资回报率,进一步提高了银行的经营利率并在一定程度上更好地为客户提供服务.这对银行拓展业务提高核心竞争力有着非常重要的现实意义.

参考文献

References

[1] 柯孔林,冯宗宪.我国商业银行效率测度及其影响因素分析[J].数理统计与管理,2008,27(1):11-16
KE Konglin, FENG Zongxian. Efficiency measurement of China's commercial banks and the determinants analysis [J]. Application of Statistics and Management, 2008, 27(1):11-16

[2] 郭迎春.知识型电力客户关系管理研究[D].保定:华北电力大学经济与管理学院,2008
GUO Yingchun. Research on knowledge-enabled customer relationship management in power enterprise [D]. Baoding: College of Economy and Management, North China Electric Power University, 2008

[3] 刘世平.数据挖掘技术及应用[M].北京:高等教育出版社,2010
LIU Shiping. Technology and application of data mining [M]. Beijing: Higher Education Press, 2010

[4] Moro S, Cortez P, Rita P. A data-driven approach to predict the success of bank telemarketing[J]. Decision Support Systems, 2014, 62(1246):22-31

[5] 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,2010:304-344
YAO Zhiyong. SAS programming and data mining for business cases [M]. Beijing: China Machine Press, 2010: 304-344

[6] 薛薇,陈欢歌.基于Clementine的数据挖掘[M].北京:中国人民大学出版社,2012:212-214
XUE Wei, CHEN Huange. Data mining based on Clementine[M]. Beijing: China Renmin University Press, 2012: 212-214

[7] 朱世武,崔巍,谢邦昌.移动电话客户流失数据挖掘[J].数理统计与管理,2005,24(1):62-68
ZHU Shiwu, CUI Wei, XIE Bangchang. Data mining on customer churn of mobile number and type [J]. Application of Statistics and Management, 2005, 24(1): 62-68

Statistical decision research for bank's long-term deposit subscription

LAI Peng[1] ZHAO Rulei[1] GUO Lizhen[1]

1 School of Mathematics & Statistics, Nanjing University of Information Science & Technology, Nanjing 210044

Abstract Nowadays, with the increasing competition between banks, it is very important to improve the bank's competitiveness by effectively excavating potential clients and providing differentiated services. The decision tree algorithm is proposed to data mine the possible 21 important attributes which affect bank clients' long-term deposit subscription. A Portuguese retail bank is addressed, with data collected from May, 2008 to November, 2010. The decision tree model is constructed to reflect the important factors in a banking client deposit subscription business. Results show that the significant factors which affect client's long-term deposit subscription are target number of bank agents, marketing duration and month. Such knowledge greatly reduces the marketing range of potential clients for term deposit thus improves the bank efficiency.

Key words data mining; customer orientation; decision tree; statistical decision