信用评分卡模型的建立

黎玉华
(华南理工大学 广东 广州 510006)

[摘 要]信用评分技术是一种应用统计模型,为信用申请者或已有的帐户计算一个风险分值的方法。而这种用途的统计模型就称为信用评分卡。信用评分卡可以根据客户提供的资料和客户的历史数据,对客户的信用进行评估。信用评分卡的建立是以对大量数据的统计结果为基础,具有相当之高的准确性和可靠性。

[关键词]特征;属性;分值

本文主要讲述了运用数据挖掘技术进行信用评分卡模型的建立原理、及建立过程分中所使用的模型和方法。建立信用评分卡模型的步骤包括:确定研究目标、确定数据源及抽取样本、数据探索、模型开发和模型验证。

1 工作原理

客户化申请评分卡是一种统计模型,它可对当前申请人的具有预测作用的申请信息进行评估并给出一个分数,该数字能定量表示申请人的预期将来表现。

客户化申请评分卡由一系列特征项组成,每个特征项相当于申请表上的一个问题(例如,供职现雇主的工作时间、银行和信用帐户参照、收入等)。每一个特征项都有一系列可能的属性,相当于每一问题的一系列可能答案(例如,对工作时间这个问题,申请人的答案有:6个月,10 年等)。在开发评分模型中,先确定属性与申请人未来信用表现之间的相互关系,然后给属性分配适当的分数权重,分配的分数权重要反映这种相互关系。分数权重越大,说明该属性表示的信用表现越好,其分数也越高。一个申请人的得分是其属性分值的简单求和。如果一个申请人的得分等于或高于用户机构所决定的界限分数,此申请人处于可接受的风险水平并将被批准。低于界限分的申请人将被拒绝,或者给予标识以便进一步审查。

2 确定研究目标

在开发信用评分卡模型的第一步,就是要清楚地确定这个模型的类型,是申请评分卡模型还是行为评分卡模型,不同的模型,由于所考虑的目标对象不同、研究的对象不一样,因此开发方法、检验手段和处理原则都不一样。本文开发的是申请评分卡模型。申请评分卡的设计目标是区分好客户和坏客户。评分卡使用结果应是:高分数的申请人意味着比低分数的申请人的风险低。

3 确定数据源及抽取样本

开发信用评分卡模型的第二步就是分析和了解所有可能使用的数据源,确定哪些数据源可以提供更加准确详细的信息。数据的类型可以分为人口统计学数据、行为数据、态度数据。信贷机构,如银行的信用卡中心,拥有足够的历史数据用于开发申请评分卡,最好能准备大约 2-2.5 年的历史交易数据。历史数据的数据量庞大,可以通过抽样是分析更具效率。

可以按照一定的处理方法和处理手段,从庞大的数据中抽取开发样本,以达到全集所具有的技术开发的潜在功能。

3.1 观测窗口与表现窗口

观测窗口表示构造模型自变量所跨越的时间段,在这个时间区间内,模型构造者收集信用卡客户的历史行为数据,它的含义是在“历史”上,给出一个“观测”区间。

表现窗口表示构造模型因变量所跨越的时间段,在这个时间区间内,模型构造者收集信用卡客户的表现数据,以此甄别“好”、“坏”客户,它的含义是在“未来”上,给出一个“表现”区间。

完备的历史数据积累对选择观测窗口和表现窗口也是一个不可或缺的因素,如果试图选择较长的观测区间,那就需要有较长时间的历史数据积累;另外数据库中一些关键变量需要要具有历史承继性以及较高的标准化程度,如果某些核心变量的稳定性随着时间的推移而发生跳跃,那么选择观测窗口和表现窗口就会局限在这些核心变量表现一致的区间。如果选择较短的表现窗口,一些需要一定时间才能表现的客户特征,如拖欠,坏帐就不能达到稳定状态。国际上通行的做法是取 6 个月、12 个月以及 18 个月等区间长度。

3.2 样本抽取

出于不同的目的,模型开发者开发出不同的行为评分,为了更好的捕捉到特定客户的行为模式,需要在采集样本时就依据不同的开发目的,采取不同的采集策略;同时由于某些特殊行为的客户,例如欺诈,坏帐,所占总样本的比例非常低,开发这些行为评分时,需要运用特殊的抽样方法,这对于全面涵盖特殊行为客户,缩短开发周期,提高建模效率都是有帮助的。

在建立申请评分卡模型时,除了考虑到样本的观察区间和表现区间外,还有样本容量的问题。样本容量应该为多少,莱温斯(Lewise)[1] 1992 年建议使用 1500 个好客户与 1500 个坏客户足够, 但在实际建模中使用到的样本数比这要大得多。在样本抽取时,应该有多少好客户、多少坏客户。如果样本中好、坏客户的数量比与总体中好、坏客户的发生比一样,就容易导致坏客户数量过少,从而不能确定坏客户的特征。所以,国际上通行的做法是,在建模时样本中好、坏客户比率为 50:50,或者位于 50:50 与总体中好、坏客户比例之间。在本信用评分卡模型建立中,使用的是逻辑回归方法,总体中好、坏客户的概率已经用于计算中,所以,当样本中好坏客户比例与总体中好坏客户的比例不一致时,我们也不需要对这一样本中得到的结果进行调整,因为回归方法中已经自动调整了。

3.3 明确数据源

对于信用卡信用评分来说包括客户的人口统计学信息和交易行为信息。信用评分卡模型的数据源通常包括如下信息:年龄、住所类型、婚姻状态、收入等等。其实这些信息,就是构成评分卡的“特征”。每一个特征又会划分为不同的组别,称为“属性”。每一个属性都有一个不同的分值相对应。评分卡还有一个临界分值(称为"临界点")来决定是否接受或拒绝一个客户的信用申请。对一个新客户申请的评估将基于该客户的个人属性。把申请客户所有属性的分值相加就得到了该客户的最后分值。再把最后分值与临界点相比较,最后分值高于临界点则表示可以接受申请该项申请,相反则将拒绝该项申请。

4 好客户和坏客户的定义

抽取的样本,包含有“好客户”和“坏客户”的样本数据,所以在做样本抽取前,必须确定如何定义“好客户”和“坏客户”。在信用评分卡模型的建立中,样本已经按照业务标准定义好坏客户。

通常坏客户的定义为: 使用信用卡历史上曾经发生过 90 天或以上拖欠。

5 数据探索

在明确了数据源和样本抽取以后,我们就开始对数据源进行数据探索。数据探索包含以下几个方面:

5.1 变量基本信息

面对包含有成百上千变量的分析型文件时,首先要了解这些变量的基本信息,如变量的属性、类型等,汇总分析出这些变量的最大值、最小值、平均值、标准差、峰值等等。

5.2 变量初选

如何选取合适的变量、将目标变量的最大程度的分离是一件十分棘手的事件。试图立即通过逐步回归方法来进行变量筛选是不太现实的,因为变量太多;同时一个重要的原因是因为某些变量会蕴含非线性的信息,变量未做变换的回归方法不能反映这一部分信息;对一些

取值为字符的变量也需要一种方法度量它对目标变量的分离度。

一般行为评分卡的变量初选的方法包括以下几种：信息价值、信息增益、卡方统计量、单变量显著水平、偏相关系数等。申请评分卡的变量初选的步骤是采用交互分组的方式筛选初步的变量。

5.3 交互式分组

在进行信用评分建模的前一步，就是对变量进行分组。交互地对“区间型”、“列名型”或“有序型”变量进行分组处理可以：

●限制变量的属性个数

●提高变量的预测能力

●选择预测变量

●增强变量之间WOE变化的平滑性和线性

在交互式分组中，我们可以对变量的分类进行交互式的修改，以便得到最佳的分组。交互式分组要求：

（1）一个二元目标变量；

（2）选择基于基尼指数（Gini score）或者信息价值（Information value）得到的显著特征项。

（3）在WOE（Weight of Evidence）值和商业考虑的基础上，对所选的特征项进行分组。

WOE是衡量一个属性相关风险的尺度。WOE的计算公式如下：

$$WOE_{attribute}=\ln(P_{attribute}^{nonevent}/P_{attribute}^{event}) \quad (1)$$

$$P_{attribute}^{event}=n_{attibute}^{event}/N^{event} \quad (2)$$

$$P_{attribute}^{nonevent}=n_{attibute}^{nonevent}/N^{nonevent} \quad (3)$$

在这公式里，N^{event} 为数据中总的坏客户的数量，$N^{nonevent}$ 为数据中总的好客户的数量，$n_{attibute}^{event}$ 为该属性里坏客户的数量，$n_{attibute}^{nonevent}$ 为该属性里坏客户的数量。一个属性的风险高低由其证据权重（Weight of Evidence）决定，WOE的值越高，说明这个分组的风险机率越低。

一个特征项能从低风险客户中分离出高风险申请者的预测能力，是由其信息价值（Information value）或基尼统计量评估出来。信息价值的值为该特征项属性的WOE的加权总和，该权值为这个属性中好客户在总好客户数中的比例与坏客户在总坏客户数中的比例的差值。具体公式如下：

$$\text{Information value}=\Sigma_{attribute}[(P_{attribute}^{nonevent}-P_{attribute}^{event})*WOE_{attribute}] \quad (4)$$

用于计算信用评分的变量的信息值应大于0.02。如果信息值大于0.5，就是过预测（over-predicting）变量。

基尼系数如下计算：

◆根据一个属性的坏客户所占比例的值，降序来排序数据，假设一个特征项有m个属性，排序的属性被分成组1、组2、…组m，每组对应一个属性，组1就是坏客户所占比例的值最高的那组。

◆对于排序后的每一组，计算组i的坏客户和好客户的数量，然后算出基尼统计量：

$$Gini=1-\left[\frac{2\Sigma_{i=2}^{m}(n_i^{event}*\Sigma_1^{i=1}n_i^{nonevent})+\Sigma_i^{m}(n_i^{event}*n_i^{nonevent})}{(N^{event}*N^{nonevent})}\right]*100 \quad (5)$$

信息价值默认的分割点为0.1，基尼系数的分割点为20。信息价值或基尼系数的值大于规定分割点的变量将被选作为评分卡的输入，否则，将给与拒绝。在交互式分组中，同时需要加入业务人员的经验，不能完全按照数据驱动方式。

6 模型建立

在信用评分卡建模中，用到的方法有很多种，其中Logistic回归是运用比较广泛的一种方法。本文信用评分建模中，我们选用的是Logistic回归模型。运用Logistic回归模型，因变量是一个二元值的概率，如坏客户或者好客户的概率。也就是说，如果我们预测一个新客户是否是默认的客户类型，用Logistic回归分析后，预测的结果并不是简单给出原始的“是”/“否”，而是给出一个这件事情将会发生的改良过的概率。这种预测的结果是连续的。Logistic回归模型的因变量取坏客户概率与好客户概率的发生比的自然对数，因为对自变量的取值没有任何限制，就不会在进行预测时出现概率小于0或大于1等此类问题，该模型提供的是坏客户的发生比，因此申请人的信用度能更准确更科学的掌握到。

在Logistic回归模型中，将概率发生比的对数表示成特征变量的线性组合，公式如下：

$$\begin{aligned}Logit(pi)&=\log(p\_good/p\_bad)\\&=\log(odds)\\&=age\_woe*b\ age+\\&\ status\_woe*b\ car+\\&+a\end{aligned} \quad (6)$$

Logit(pi)是一个对数值，即为log(P(bad)/P(good))，由于P(bad)/P(good)的取值在0到∞，log(P(bad)/P(good))的取值在-∞到+∞之间。

为了使获得的评分更具“实用性”，需要对每个属性的分值需要进行线性比例变换，然后再加上一个偏移量。评分和用于逻辑回归（Logistic Regression）建模的好/坏比（good/bad odds）的对数成比例，而不是好/坏比（good/bad odds）本身。所以分值可以是负数，而且越小的分值代表风险越高。

每个属性对应的分值可以通过下面的公式计算：WOE乘该变量的回归系数，加上回归截距，再乘上比例因子，最后加上偏移量：

$$(woe_i*\beta_i+\frac{a}{n})*factor+\frac{offset}{n} \quad (7)$$

对于评分卡的分值，我们可以这样计算：

$$\begin{aligned}score&=\log(odds)*factor+offset\\&=(\sum_{i=1}^{n}(woe_i*\beta_i)+a)*factor+offset\\&=(\sum_{i=1}^{n}(woe_i*\beta_i+\frac{a}{n}))*factor+offset\\&=\sum_{i=1}^{n}((woe_i*\beta_i+\frac{a}{n})*factor+\frac{offset}{n})\end{aligned} \quad (8)$$

比例因子和偏移量的确定方法如下：

●令好/坏比=50/1对应的评分为600

●在此基础上评分值增加20可以使好/坏比翻番

所以：

$$600=\log(50)*factor+offset$$

$$620=\log(100)*factor+offset$$

$$factor=20/\log(2)$$

$$offset=600-factor*\log(50)$$

7 模型验证

模型验证是十分重要的一步。模型开发以后，该评分卡的预测性非常好，而且也可以把工程开发过程中揭示的用户商务目标考虑进去。这种评分卡利用部分样本数据（通常训练样本的80%）研制而成，并利用剩余的样本数据（通常占开发样本的20%）进行验证。在模型正式建立以前（包括模型初步拟和），不得动用测试样本；整个模型建立过程中，不得动用校验样本。

通过对比训练样本和校验样本的各种统计量，如客户的最大和最小的分数、KS值等，验证模型是否可以推广应用。科

【参考文献】

[1]林功实，林健武.信用卡.北京：清华大学出版社，2006.

[责任编辑：王静]

（上接第463页）以实现我们所需要的功能，而这个对象的具体方法做了什么，具体过程，我们就不用去关注。这就是面向对象程序设计的思想。科

【参考文献】

[1]雍俊海. Java程序设计教程[M].北京：清华大学出版社，2007.

[2]陈月峰.浅谈Java面向对象程序设计[J].电脑知识与技术，2009(33).

[3]马鲁宁.JAVA语言面向对象程序设计的特点[J].黑龙江科技信息，2007(2).

[4]笪斌，苏建元.实现面向对象的继承性封装性和多态性[J].电脑学习，2005(6).

作者简介：黄俊爽（1987—），女，汉族，河南南阳人，河南师范大学，计算机与信息技术学院。

[责任编辑：张艳芳]