DOI:10.3969/j.issn.1003-1154.2017.02.001

基于GBDT与Logistic回归融合的个人信贷风险评估模型及实证分析

□ 蔡文学 罗永豪 张冠湘 钟慧玲

(华南理工大学 经济与贸易学院，广东 广州 510006)

[摘要] 随着个人信贷业务的快速发展，如何评估个人信贷风险是一个重要的问题，通过GBDT模型从原始数据中提取组合特征，再使用Logistic回归构建个人信贷风险评估模型，最后对个人的信贷数据进行实证分析，可知GBDT与Logistic回归融合模型在与其他模型相比具有更高的信贷风险预测准确性，"信贷情况良好"的预测准确率达到了87.6%，"信贷情况不良"的预测准确率达到了81.3%。

[关键词] 个人信贷风险评估；GBDT；组合特征；Logistic回归

[中图分类号]F832.332 [文献标识码]A [文章编号]1003-1154(2017)02-0001-04

一、文献综述

金融创新和个人消费观念的改变促进了国内银行信贷业务特别是个人信贷业务的快速发展。然而贷款业务规模的扩大的同时，也给银行带来了巨大的风险，若借贷人不能或不愿按照信贷协议约定偿还本息，就会有对银行的经营造成损失，如何在业务扩展的同时，有效控制和防范个人信贷风险是各大银行亟待解决的重要问题。因此，构建个人信贷风险评估系统，及时发现危险的贷款申请，有效地应对可能发生的个人信贷风险，不论对商业银行自身而言，还是对监管机构而言，都具有重要的现实意义。

个人信贷风险评估问题一个重要的解决途径，是根据过去个人的信贷信息以及是否违约的记录建立信贷风险评估模型，预测用户是否会出现违约的情况。发达国家的商业银行有着比较成熟的个人信用评估系统，但是我国很多商业银行的信用评估体系仍然处于摸索的阶段，即使已经构建的也存在不少问题。总的来说，我国的信用评估体系主要存在两个方面问题，一是部分商业银行照搬发达国家的信用评估模型，但是由于我国与其他国家在体制和文化上的差异，生搬硬套的模型会带来许多问题；二是指标选取和特征处理的问题。影响个人信贷信用评估的因素较多，包括用户的年龄、性别、学历、职业以及信用卡消费、贷款等行为，使用简单的特征处理和选择方法可能会没有充分利用好原始的数据，忽略掉部分重要的信息，出现预测结果不理想的情况。

因此，以基础的业务为基础，充分利用历史数据，提出合适的个人信贷信用评估模型，能够有效解决上述问题，提高预测和评估的准确性。

纵观目前关于个人信贷评估模型的研究，主要的方法包括BP神经网络[1]、SVM[2]、层次分析法[3]、Logistic[4-6]回归等。其中，Logistic回归模型在个人信贷评估问题上是一种实用性很强的模型，陈鹿婧等[4]以违约的概率作为信用评估风险衡量标准，构建P2P机构的借款人信贷风险的Logistic模型，通过实证证明logistic回归模型具有较高的准确性。张国政等[5]基于商业银行的个人消费信贷的实际数据，使用Logistic回归构建个人信贷评估模型，通过实证测得了影响个人信贷风险的关键因素。廖绚和李兴绪[6]依据5P原则及其他因素，利用Logistic构建个人信贷风险评估模型，对各个因素于违约之间的相关程度进行了实证分析，并对银行信贷风险进行了评估，实证表明逻辑回归是一种高性能的评估模型。

综上所述，Logistic回归在个人信贷风险评估问题上具有较高的准确性，是一种常用的预测模型。然而，Logistic回归是一种线性模型，本身的学习能力有限，对特征处理的要求比较高，如果特征处理不当，模型会出现重大的缺陷，主要体现在两个方面：一是如果特征变量过多，会导致Logistic回归模型出现多重共线性的问题，降低预测准确率；二是如果特征的解

[基金项目] 国家社会科学基金项目(14BGL139)；广州市科技计划项目(201510010194)；中央高校基本科研业务费资助项目(2015ZDXM06)。

释性不足，特别是当数据量较大和特征变量较少时，采用 Logistic 回归模型往往得不到理想的效果。针对第二种情况，一种解决方案是在原有特征的基础上增加组合特征，通过扩展数据的维度间接提高模型的学习能力，再在此基础上建立 Logistic 回归模型，从而提高模型的预测精度。组合特征的提取可以使用机器学习的方法，Facebook 等[7]于 2014 年提出将 GBDT 和 LR 的融合模型应用于 CTR 预估问题上，采用 GBDT 模型提取组合特征，再使用 Logistic 回归建立 CTR 预估模型，成功地提高了预测的准确率。而本文则将 GBDT 与 LR 的融合模型应用于个人信贷风险评估中。

二、模型理论基础

（一）GBDT 模型构造组合特征

GBDT（Gradient Boost Decision Tree）是一种常用的非线性模型，基于 boosting 算法的思想，每次迭代都在减少残差的梯度方向新建立一棵决策树，通过迭代不断提高预测的准确性。由于 GBDT 能够发现多种有区分性的特征以及特征组合，决策树的路径可以直接作为其他模型的输入特征使用，省去了人工寻找特征、特征组合的步骤。

GBDT 是决策树的组合模型，使用 GBDT 构造组合特征是指将 GBDT 中所有决策树每一个叶子节点作为一个新的特征，因此构造得到的特征数目与 GBDT 叶子节点的数目相同，每一个特征取值为 0 或者 1。对于每一棵决策树，若输入的样本落入到某个叶子节点，则该叶子节点的取值为 1，否则为 0。设 GBDT 模型中决策树集合为 $T=\{T_1,T_2,\cdots,T_m\}$，其中 T_i 表示第 i 棵决策树，设 $l_i=\{l_{i1},l_{i2},\cdots,l_{ik}\}$ 表示样本 S 在决策树 T_i 各叶子节点上的取值，$l_{ij}\in\{0,1\}$，若样本 S 输入到决策树 T_i 并且落入到叶子节点 j 中，则 $l_{ij}=1$，否则 $l_{ij}=0$，将样本 S 输入到 GBDT 中，得到该样本对应的新的组合特征数据 $l=\{l_1,l_2,l_3,\cdots,l_m\}$，$m$ 为 GBDT 所有决策树叶子节点数目之和。

图 1 中，通过对原始数据训练得到 GBDT 模型，模型包含两棵决策树 Tree1 和 Tree2，共 5 个叶结点，每一个叶结点作为一个新的特征，若特征落入叶结点中，则该特征标志为 1，否则为 0。样本 X 经过 Tree1 落入第一个叶子节点，经过 Tree2 落入第 2 个叶子节点，则新构造的特征向量为[1,0,0,0,1]。

GBDT 具有两个重要的参数，一是每棵决策树最大叶子节点数目，本文使用 α 表示，二是决策树的数目，本文使用 β 表示。α 和 β 的取值决定了组合特征的数目，如果 α 和 β 取值过大，组合特征数目过多，会出现特征冗余、过拟合的情况，如果 α 和 β 取值过小，

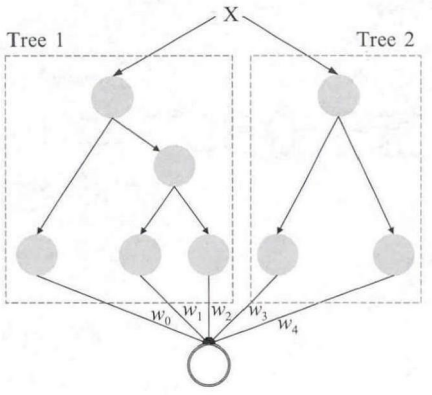

图 1　GBDT 构造特征示意图

组合特征数目过少，对 Logistic 回归模型的改进难以体现。选取合适的参数是 Logistic+GBDT 模型实现的重要环节，本文将在实证分析中通过实验选取最优的 α 和 β。

（二）GBDT 与 Logistic 融合模型

Logistic 回归是一种基本的二分类模型，在线性回归的基础上通过 sigmoid 函数将输入函数值映射到 0～1 区间，表示，作为各类判别的概率。设 y 为一个二类的因变量，表示个人的信贷信用情况，$y\in\{0,1\}$，其中 $y=1$ 表示个人信贷信用不良，$y=0$ 表示个人信贷信用良好，$x=\{x_1,x_2,x_3,\cdots,x_p\}$ 为相应的 p 维解释变量。概率 $p(y=1|x,\theta)$ 表示在给定特征向量 x 的条件下 y 属于类别 1 的概率，令 $h_\theta(x)=p(y=1|x,\theta)$，$h_\theta(x)$ 可以表示如下：

$$h_\theta(x)=\frac{1}{1+e^{-\theta^T x}} \tag{1}$$

称（1）为逻辑回归模型，其中 $\theta=\{\theta_0,\theta_1,\cdots,\theta_p\}$ 表示各个特征对应的系数。

参数 θ 可通过求解最大似然估计函数获得，由于 y 为二类变量，y 在给定 x 的条件下服从两点分布，则似然函数为：

$$l(\theta)=\prod_{i=1}^{n} h_\theta(x)^{y_i}\cdot(1-h_\theta(x))^{1-y_i} \tag{2}$$

通过梯度下降等参数估计方法可以求得参数 θ。

本文将 GBDT 与 *logistic* 回归进行融合，通过 GBDT 模型构造新的组合特征，然后与原始特征结合训练 Logistic 模型，则 GBDT 与 Logistic 融合模型的训练和预测过程如下：

步骤 1：设置参数 α 和 β，通过原始数据训练 GBDT 模型，用于构造组合特征。

步骤 2：对于每一个训练样本，输入到步骤 1 训练得到的 GBDT 模型，每一个叶子节点的输出组成组合特征向量。

步骤 3：步骤 2 中得到所有样本的组合特征向量与原始特征向量进行合并，得到新的数据样本，用新

的数据样本训练得到的 Logistic 模型作为最终的预测模型。

步骤 4：对于样本的预测，首先将预测样本输入到 GBDT 模型中得到组合特征向量，然后将组合特征向量与原始特征向量合并成新的特征向量，输入到 Logistic 模型中，得到的输出即为该样本的预测结果。

三、实证分析

（一）数据来源

本文使用的数据来自于融 360 提供的用户贷款数据，包括 2.5 万用户的基本属性、银行流水、信用卡账单记录、浏览行为、放款时间等数据。本文采用违约情况来衡量个人信贷风险，若用户出现过违约情况则认为该用户“信贷情况不良”，没有出现过违约情况的用户是“信贷情况良好”，原始数据中共6 783名用户标志为“信贷情况不良”，18 595名用户标志为“信贷情况良好”。为了提高实证的准确性，本文采用多次随机实验进行数据验证，将原始数据集划分为训练集和测试集，共进行 60 次实验，训练集和数据集的划分情况如表 1 所示。

表 1　原始数据划分

样本总量	训练集	测试集	比例	次数
25 378	22 840	2 538	9∶1	20
25 378	20 302	5 076	8∶2	20
25 378	17 765	7 613	7∶3	20

（二）特征选取

本文对原始数据进行特征提取，从基本属性、信用卡信息、信用卡消费行为、账单信息 4 个方面选取了 20 个与用户信贷风险评估相关的特征变量，如表 2 所示，共包括性别、年龄、职业、婚姻状况、户口类别、个人月收入、信用卡张数、使用频率、银行卡跨行数目、信用卡额度、月刷卡次数、月刷卡金额、最大月刷卡次数、最高月刷卡金额、月还款额少于应还金额的次数、信用卡可用余额。

表 2　基础特征

特征名称	符号	特征名称	符号
性别	X1	信用卡张数	X11
年龄	X2	使用频率	X12
职业	X3	信用卡跨行数目	X13
婚姻状况	X4	信用卡额度	X14
户口类别	X5	月均刷卡次数	X15
个人月收入较低(低于 1 000)	X6	月均刷卡金额	X16
个人月收入较低(1 000～5 000)	X7	最大月刷卡次数	X17
个人月收入一般(5 000～1 000)	X8	最高月刷卡金额	X18
个人月收入较高(10 000～20 000)	X9	月还款额少于应还金额的次数	X19
个人月收入很高(20 000 以上)	X10	信用卡可用余额	X20

上述特征变量中有 9 项是离散型变量，其余各项是连续型变量，由于逻辑回归是一种基于线性回归的模型，因此在进行最终逻辑回归参数训练数需要对特征向量进行标准化。对于离散型变量，采用特征编码的方式进行处理，使用数值 1，2，3…进行表示，对于连续性变量，采用 Min Max 方法进行标准化，如公式(3)所示，将数据映射到 0～1 之间。

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3}$$

其中，x' 表示经过标准化后的数值，x 表示原始数值，x_{min} 表示变量的最小值，x_{max} 表示变量的最大值。

（三）实验结果分析

1. GBDT 参数设置

本小节通过数据实验选取最佳的参数 α（每棵决策树最大叶子节点数目）和 β（决策树数目），采用总体预测准确率的平均值作为实验的评价。一共进行两组实验，第一组实验用于确定参数 α 的值，首先固定 β 的取值为 20，将原始数据通过 GBDT 生成新的组合特征，再将新的特征与原始特征组成新的特征变量训练 Logistic 模型，得到参数 α 与预测准确率的关系如图 2 所示，纵坐标为预测准确率的变化情况，横坐标为 α 的取值，准确率经过一个上升再逐渐平稳的过程，当参数 α 为 9 时准确率的曲线开始平稳，10～20 的准确率上升幅度低于 0.1%，因此本文选参数 α 的值为 9。第二组实验用于确定参数 β 的值，固定 α 的取值为 9，通过实验测得参数 β 与预测准确率的关系如图 3 所示，当 β 处于开始阶段时，准确率上升速度较快，然后逐渐减弱，当 β 处于 15～50 区间时，上升趋势逐渐趋于平稳，当 β 超过 25，准确率几乎没有提升，因此本文选择的参数 β 为 25。

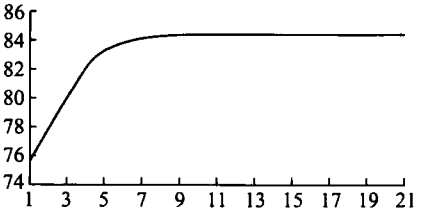

图 2　α 取值与预测准确率之间的关系

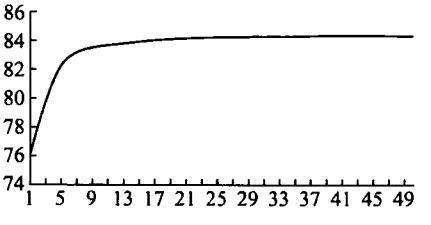

图 3　β 取值与预测准确率之间的关系

2. 对比实验

本节采用多次随机实验预测准确率的平均值以及方差衡量模型的性能，在每次实验中，使用事先划分好的训练数据集进行模型的训练，然后使用测试数

据集对模型进行预测，预测正确的样本数量与预测样本的总数量的比例即为该次实验的预测准确率，其平均值反映了预测模型的预测精度，方差反映了其预测的稳定性。

设置参数 α 和 β 分别为 9 和 25，使用原始数据训练 GBDT 模型，模型所有叶子节点的取值作为一组特征与原始特征构成新的特征向量，然后使用新的特征数据训练 Logistic 模型，得到 GBDT 与 Logistic 融合的个人信贷风险评估模型，本节使用 GBDT＋Logistic 表示。为了验证模型的有效性，本节将使用第一节提及的训练集进行多个模型对比实验，其他的对比模型包括：(1)随机森林：决策树数目 200；CART 中的分割属性选择 Gini 指数；(2)GBDT：决策树数目 1 000；(3)SVM：径向基核函数，惩罚系数 C＝1；(4)朴素贝叶斯；(5)BP 神经网络(通过多次实验调整参数发现隐藏层数目为 40 以及迭代次数为 218 较优并趋于稳定，故选择固定参数 40 和 218)；(6)Logistic 回归：L1 正则化、连续特征离散化。60 次实验的结果如表 3 和表 4 所示。

表 3 “信贷情况良好”预测结果对比

模型	预测准确率			方差
	最低	平均	最高	
随机森林	83.4%	85.4%	88.1%	1.92
SVM	81.3%	83.7%	86.2%	3.02
GBDT	83.5%	86.2%	90.2%	3.21
朴素贝叶斯	76.4%	80.3%	84.5%	3.62
BP 神经网络	78.4%	80.8%	81.7%	0.91
Logistic 回归	79.1%	82.1%	85.6%	2.49
GBDT＋Logistic	85.2%	87.6%	89.8%	1.87

表 4 “信贷情况不良”预测结果对比

模型	预测准确率			方差
	最低	平均	最高	
随机森林	74.8%	76.7%	79.4%	2.03
SVM	72.8%	75.4%	78.9%	2.93
GBDT	73.6%	77.8%	80.8%	3.35
朴素贝叶斯	66.5%	70.2%	73.9%	3.09
BP 神经网络	70.2%	71.3%	72.7%	1.05
Logistic 回归	70.8%	73.9%	76.9%	2.94
GBDT＋Logistic	78.7%	81.3%	83.4%	1.95

从”信贷情况良好”预测准确率上看，所有模型的平均预测准确率都超过了 80%，表明在“信达情况良好”的预测上，所有模型的预测准确率比较高，其中 GBDT＋Logistic 模型效果最佳，预测准确率达到了 87.6%，相比单一 Logistic 回归模型提高超过了 5%，但是随机森林以及 GBDT 两个组合模型的准确率也超过 85%。从”信贷情况不良”预测准确率上看，朴素贝叶斯预测效果最差，预测准确率只有 70.2%，其次是 BP 神经网络，单一模型中最好的是 SVM，准确率为 75.4%，而 GBDT＋Logistic 模型的预测准确率达到了 81.3%，远高于组合模型随机森林的 76.7%以及 GBDT 的 77.8%，比单一 Logistic 回归模型提高超过了 7%，表明 GBDT＋Logistic 模型在预测“信贷情况不良”时性能明显优于其他算法。从稳定性上看，无论在“信贷情况良好”还是“信贷情况不良”的预测上，GBDT＋Logistic 模型的方差仅高于 BP 神经网络，具有较高的稳定性。综上，在个人信贷风险评估问题上，与其他算法相比，GBDT＋Logistic 模型能够获得更优的预测效果。

四、结　论

本文将 GBDT 与 Logistic 回归模型融合应用于个人信贷风险评估问题上，并以融 360 提供的用户贷款数据进行实证分析，结果表明 GBDT 与 Logistic 回归模型与其他模型相比具有更高的风险预测准确率和稳定性。Logistic 回归模型是一种线性分类器，自身的学习能力有限，通过 GBDT 模型从原始数据中提取有效的组合特征，能够预先分析出有效的特征、特征组合，充分利用历史数据，避免忽略原始数据中的重要信息，从而提高预测的精度，实证结果证明 GBDT＋Logistic 模型的预测效果远优于单一的 Logistic 回归模型。银行信贷数据具有指标多、噪声复杂的特点，提出使用 GBDT 与 Logistic 回归融合的个人信贷风险评估模型，对实际应用具有重要的指导意义。□

[参考文献]

[1] 胡绪华，吉敏. 基于 BP 神经网络的银行信贷风险评价[J]. 统计与决策，2009(11).

[2] 张奇，胡蓝艺，王珏. 基于 Logit 与 SVM 的银行业信用风险预警模型研究[J]. 系统工程理论与实践，2015(07).

[3] 陶涛. 基于层次分析法的银行信贷资产证券化风险控制分析[J]. 统计与决策，2013(06).

[4] 陈鹿婧，杨青骥，孙超凡，汪小燕. 基于 logit 模型的 P2P 公司的个人信贷风险评估[J]. 新经济，2016(20).

[5] 张国政，姚珍，杨亦民. 基于 Logistic 回归的农户小额信贷风险评估实证研究[J]. 财会月刊，2016(27).

[6] 廖绚，李兴绪. 基于 Logit 模型的银行个人信贷风险管理评估[J]. 统计与决策，2008(21).

[7] He X, Pan J, Jin O, et al. Practical Lessons from Predicting Clicks on Ads at Facebook[M]. ACM, 2014.