基于分类模型对个人信用风险评估的研究

顾佳青[1]　陈涛[2]
（南昌大学统计学，江西　南昌　330000）

【摘　要】信用风险属于一种行为风险，产生的原因有客观（履约能力问题）和主观（履约意愿问题）两方面，而目前大部分消费金融主要业务是围绕个人信贷来开展，因此如何正确评估个人信用风险是最重要的关键所在。本文首先构建大数据征信指标体系，提出了将极端梯度提升算法的输出结果进行独热编码转化，和筛选出入模的特征变量重新组合作为逻辑回归模型的输入变量，形成极端梯度提升算法加逻辑回归两阶段的信用评估模型，研究结果表明逻辑回归模型的稳定性和准确率最高。同时指出当极端梯度提升的预测分类效果达到较优时，融合逻辑回归模型在分类预测准确性方面有时反而偏弱，为金融机构在个人信贷业务的数据获取及模型选择上提供了一定的参考价值。

【关键词】大数据征信；极端梯度提升；逻辑回归；独热编码

一、引言

国外欧美发达国家是最早开始提出信用消费理念的，而我国是1985年才出现了第一张信用卡。随着国内居民生活水平的提高，老百姓开始逐渐意识到自身良好的信用是可用以提高生活的品质。据此信贷需求规模开始逐渐扩大，但实际我国个人征信体系的发展却未完全跟上。随着智能手机的普及，提供信贷产品的模式移到线上使得互联网金融就突然的崛起[1]，传统金融无法获取消费信贷用户，因为互联网消费金融的出现而得到很好的服务。目前传统金融仍旧是基于人行数据的征信模式在进行风险评估，无论是数据量、维度、时效都无法跟上互金时代的步伐，如何避免产生较大的信用风险也是传统金融面临急需解决的问题。而非传统金融机构由于没有对风控足够的敬畏因而欺诈和坏账也不断攀升。除了风险管理体系外还有数据整合与金融行业经验的欠缺，甚至有些仅靠基础信息审核是存在较大隐患的。而大数据征信是用户基于互联网场景产生的实时行为数据，恰巧能作为传统数据有效的补充，同时解决了实时性和真实性的问题。因此结合互金征信与传统人行的信用风险评估研究是完全顺应大数据时代的需求。

本文收集脱敏的个人信用贷款数据，结合人行征信及大数据征信，以信贷是否存在违约行为作为切入点，通过对行业内普遍实操采用的分类模型的分析，提出将非线性分类模型输出变量转化并导入到线性模型中训练的融合评估模型，希望为金融机构今后在选择模型上提供更为客观的理论基础。

二、信用风险评估理论分析

目前国内银行采用传统的信用评估方法通常基于人行数据，所关注的因素太少评估准确性较低。其他数据收集难度大、效率低下，难以满足互金的体验。本文使用大数据征信建立的指标体系，通过对比选择更合适的算法组合来构建信用评估模型，再引入信贷领域常用的模型评价指标进行评价。

（一）信用风险评估主要方法比较和选取

从评估方法的复杂程度来区分模型可以分为线性与非线性模型，它们的主要区别在于模型的假设和数学描述不同。文献[2]通过对多种分类模型在信用风险评估的对比发现，线性模型的优点是简单易懂，容易实现，并且可以对一定规范内的数据进行准确预测，而非线性模型的优点是可以更好地穿透复杂数据之间的关系，预测准确性就会相应提高。通过对比，本文选取传统金融常用的Logistic回归和Xgboost模型。

（二）模型评价指标

信贷领域中，由于业务上常以评分形式进行风险预测结果的输出，所以通常使用AUC和KS两类指标进行模型评价[3]。其中，KS值是用来比较两组数据的差异，以探索两者是否具有显著性差异[4]。通常业界认为：当KS值<0.2时，模型几乎没有预测能力，当KS值>0.3则基本认为模型具有一定的预测能力，行业普

遍认同是在0.4-0.5之间属于最优且合理预判。

三、个人信用风险指标体系搭建

笔者根据获取的信息最终分类出三大模块数据，分别是传统指标、人行征信以及大数据征信指标。大数据征信构成的指标举例如下：传统金融含保险、贷款、个人/对公业务、银行等，互联网金融含消费平台、消费金融公司、互联网金融门户等，还有电商信息等。

四、搭建个人信用风险评估模型

构建完成了指标体系工作后，就进入建模环节，先通过特征工程处理挑选合理入模的最终变量，再利用Logistic回归和Xgboost两种算法，分别建立风险评估模型并对模型的效果进行检验和评价。

（一）数据预处理

通常真实情况不是所有申请用户都有信贷记录，因此数据往往是不完整、含有噪声的，因此特征工程的首先就是对数据做预处理，该步骤对于数据挖掘及后续模型的搭建尤为重要。采用行业内成熟的处理方案：先进行缺失值的删补，然后进行数据去重和标准化（去除异常值和统一格式转换）。

（二）特征变量处理

本文的样本特征变量较多，复杂的统计方法可能计算量较大且效果不明显，本文采用信贷风险分析领域中，最常用的是对连续性或离散型变量进行WOE转换，然后再进行IV值的换算，计算信息价值贡献，通过IV值的大小筛选变量。主要筛选IV值>0.1且变量与变量之间相关系数<0.7的变量，筛选出总共64个变量。

（三）构建个人信用评估模型

本文模型检验不是单一模型上进行参数比较，而是改用正反向对比来验证模型效果，反向模型KS代表模型最高可能达到的结果。如果反向训练集效果过差，则说明测试集数据本身分布较为特殊，需要重新划分数据。

LR模型分类效果：模型在训练集上的KS为0.461，AUC为0.796，在测试集上KS为0.413，AUC为0.771。通常风控模型的KS在0-0.6之间，KS超过0.6一般都存在过拟合现象，并且在行业的建模实操中，普遍公认KS值达到0.3模型即为可用，本次训练模型KS在训练集及测试集上均在0.4以上，已经完全满足预测的标准。XGB模型分类效果：模型在训练集上的KS为0.541，AUC为0.843，在测试集上KS为0.413，AUC为0.766。通过与逻辑回归模型的分类效果对比，虽然训练集上的表现比逻辑回归更优，但是测试集上表现没有太大区别，从训练集和测试集的KS差距较大可以看出，模型稳定性对比逻辑回归更波动，这也正好符合之前我们对两者模型的描述比较，Xgboost确实对大数据征信类型有更优的效果，但是稳定性主要还是逻辑回归会更占优势。

五、组合模型和单一模型比较

此次组合模型主要是使用Xgboost与Logistic模型进行融合，其实早在2014年就有相关学术论文[5]介绍了通过梯度提升决策树结合逻辑回归处理的方案。此次模型对比构建的流程如下：

1. 首先要对样本进行训练集和测试集切分（按时间线切分）；

2. 然后开始训练xgboost模型，并根据训练后的模型去直接预测表现；

3. 试算训练集和测试集分别得出的KS值和AUC值进行比较；

4. 通过apply()接口获取拟合后训练集和测试集的叶子结点索引；

5. 调用sklearn.preprocessing中的OneHotEncoder进行转化向量；

6. 然后将转换后的向量的训练集，放入到LR模型中进行训练并预测；

7. 训练后的模型去直接预测测试集的表现，分别得出的KS值和AUC值进行比较；

8. 然后再将原来训练集和测试集，与通过编码后的训练集和测试集，两套样本进行合并成新的样本；

9. 再次放入LR模型中进行训练，并预测得出的KS值和AUC值进行比较。汇总各类结果如表5-1所示。

表5-1 各类样本及模型预测

模型及样本	结论值
xgboost train auc	0.84335
xgboost test auc	0.76663
xgboost train ks	0.54090
xgboost test ks	0.41328
基于Xgb特征编码后的LR train auc	0.89551
基于Xgb特征编码后的LR test auc	0.7434
基于Xgb特征编码后的LR train ks	0.41328
基于Xgb特征编码后的LR test ks	0.37365
基于组合特征的LR train AUC	0.89851
基于组合特征的LR test AUC	0.7418
基于组合特征的LR train ks	0.63448
基于组合特征的LR test ks	0.37622

通过对比发现融合 xgboost 特征编码后的模型预测结果，训练集的准确性确实比单一 xgboost 模型预测来的更高，但是区分度上反而变弱了，而且在测试集上的准确性以及区分度上都有一定程度变弱，将特征编码与原始数据融合后再次通过 LR 模型预测，只是效果上略微比使用特征编码的模型好一点点，但是仍旧没有单一 xgboost 模型预测的效果好，可见 xgboost 在此次模型对比中是最强的也是相对稳定性更高的。

考虑到可能其他梯度决策树模型的分类能力会更高，或者使用相同构建流程会找到更适合的组合模型，因此将 Random Forest 以及 GBDT 模型共同纳入进来进行对比，包括进行融合 LR 模型以及单一模型效果同时做比较。

表 5-2 各模型获取的预测结果

模型对测试样本进行预测	结论值
RF+LR 的 AUC 为	0.748624162
RF 的 AUC 为	0.744016248
GBT+LR 的 AUC 为	0.742544084
GBT 的 AUC 为	0.745763137
xgboost+LR 的 AUC 为	0.730618107
xgboost 新特征与原始特征 +LR 的 AUC 为	0.732662329
xgboost 的 AUC 为	0.765721512
Logistic 的 AUC 为	0.771321434

首先，对于所有模型测试集表现来看，LR 的模型仍旧是最佳。这也正好反映了为何目前大部分金融机构主流的还是使用逻辑回归模型，其次，对比随机森林和 GBDT 分别与 LR 进行融合后的结果，与单独使用随机森林和 GBDT 模型预测的结果相差不大，说明两者融合效果相对有限。其次，集中围绕 Xgboost 观察，Xgboost 单一模型的效果最优，转换特征 + 原始特征后训练 LR 模型略微比特征转化后训练 LR 模型提升一点点。这与 Kaggle 上的部分模型比赛实操结果有很大的不同，也与笔者的预期不符。

六、结语

本文主要是在个人信用风险评估问题方面的分析展开，主要研究的是如何构建传统数据与大数据征信相结合的指标体系，并在此基础上完成各种分类模型对个人信用风险评估的理论研究与实证分析，根据获取的信贷数据观察组合模型与单一模型预测能力的对比情况。

从最终的效果来看，单一模型预测效果最佳的还是逻辑回归模型，说明目前金融信贷行业普遍采用逻辑回归作为信用评分主模型还是经得起推敲的。因为作为金融行业对风险评估的要求，稳定性始终是第一考量。从此次研究分类模型预测值的对比来看，由 Xgboost 和 Logistic 回归模型构建的两阶段组合模型在预测准确率上反而比单一模型略微偏弱，说明单纯从算法分析并非所有场景的样本上实践都有优势，从模型稳定性来看，也是 Logistic 回归模型略优于 Xgboost 模型，所以总体而言 Xgboost-Logistic 回归两阶段组合模型效果还是受样本分布及业务属性制约，对金融机构采纳何种分类模型进行逾期预测具有一定的参考价值。

参考文献：

[1]A. Fensterstock. The Internet: The Future of Credit [J]. Journal of Shanghai Jiaotong University, 2000, 5(1): 225-229

[2] 石庆焱 . 一个基于神经网络 -Logistic 回归的混合两阶段个人信用评分模型研究 [J]. 统计研究 , 2005, (05): 45-49

[3] 黄秋彧 , 史小康 . 个人信用风险评分模型的评价指标研究 [J]. 新疆财经 , 2015, (04): 5-13

[4] 于晓阳 . 互联网 + 大数据模式下的征信—以芝麻信用为例 [J]. 北方金融 , 2016, (11): 73-76

[5]X. He, J. Pan, O. Jin, et al. Practical lessons from predicting clicks on ads at facebook[C]. Proceedings of 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, 2014: 1-9

作者简介：

顾佳青（1984.8-），男，汉族，上海市，硕士研究生，本科，就读南昌大学统计学，金融统计模型研究；

陈涛（1963.9-），男，汉族，江西省南昌，硕士生导师，教授，博士。