IMI No. 1925
Working Paper

消费金融大数据信用评估研究
——基于 GBDT、XTBoost

戴稳胜　刘志兴

【摘　　要】内需逐步成为我国经济发展的重要支柱，发展消费金融、促进消费升级成为近年小微金融发展的一个方面随之而来的过度授信及带来的违约问题也成为消费金融领域较为严重的问题。完善信用评估方法、以严谨的信用评估结果为依据，将违约风险较高客户阻挡于授信之前是消费金融风险防范的重要手段。本文利用 M 信托消费金融业务真实放款的历史数据（脱敏）为建模样本，结合第三方数据源的大数据优势，应用基于某信托机构消费金融数据，构建了基于大数据的机器学习信用评分模型，应用 GBDT 算法与 XGBoost 算法进行信用评分建模，并将结果与传统的 Logistic 建模结果进行比较，结果发现，无论是在数据建模区分结果的 K-S、AUC 指标及稳定性 PSI 指标上，还是在实际经济效益的拒绝率与违约率的表现上，基于 GBDT 算法与 XGBoost 算法的模型表现均优于传统 Logistic 模型。

【关　键　词】消费金融；信用评估；大数据建模；GBDT 算法；XGBoost 算法

【文章编号】IMI Working Paper No.1925

消费金融大数据信用评估研究

——基于GBDT、XTBoost

戴稳胜[1]　刘志兴[2]

【摘要】内需逐步成为我国经济发展的重要支柱，发展消费金融、促进消费升级成为近年小微金融发展的一个方面随之而来的过度授信及带来的违约问题也成为消费金融领域较为严重的问题。完善信用评估方法、以严谨的信用评估结果为依据，将违约风险较高客户阻挡于授信之前是消费金融风险防范的重要手段。本文利用 M 信托消费金融业务真实放款的历史数据（脱敏）为建模样本，结合第三方数据源的大数据优势，应用基于某信托机构消费金融数据，构建了基于大数据的机器学习信用评分模型，应用 GBDT 算法与 XGBoost 算法进行信用评分建模，并将结果与传统的 Logistic 建模结果进行比较，结果发现，无论是在数据建模区分结果的 K-S、AUC 指标及稳定性 PSI 指标上，还是在实际经济效益的拒绝率与违约率的表现上，基于 GBDT 算法与 XGBoost 算法的模型表现均优于传统 Logistic 模型。

【关健词】消费金融;信用评估;大数据建模;GBDT 算法;XGBoost 算法

一. 前言

2009-2018 年期间，作为拉动经济增长的三驾马车，消费对 GDP 的增长贡献率逐年递增，国家统计局公布的最近 10 年宏观经济数据显示，2009 年度消费支出对经济增长的贡献率还仅仅为 56.1%，2018 年度就达到了 76.1%，随着国内外形势的变化，高质量内需可能成为经济发展的最重要支柱，在“三驾马车”中，消费的“压舱石”作用愈发明显。在这样的背景下，消费金融也得到长足发展，市场参与者除了各家消费金融公司积极布局外，电商公司、网络小贷以及包括阿里、腾讯、京东、苏宁等互联网巨头也纷纷介入消费金融市场，市场规模也不同扩大。以 2015 年 4 月上线的蚂蚁花呗，截至 2015 年年末，蚂蚁花呗的累计放贷规模就已超过 800 亿元。但与之相伴的就是，消费金融也是乱象丛生，违规校园贷、高利贷、套路贷、暴力催收、P2P 爆雷等一系列问题频发。为规范消费金融行业，监管部门

[1] 戴稳胜，中国人民大学国际货币研究所研究员、中国人民大学财政金融学院教授
[2] 刘志兴，中国人民大学财政金融学院

于 2017 年 6 月发布了《关于进一步加强校园贷规范管理工作的通知》，2017 年 12 月发布了《关于规范整顿“现金贷”业务的通知》，正式开启了监管部门持续规范消费金融行业的严监管模式。为顺应形势规范经营，避免防范信用风险过程中出现不规范行为，消费金融机构理应加强信用风险评估，力争将消费金融信用风险消灭在业务流程之初。同时，消费金融与传统金融业务有很大不同，具有金额小、业务额度小、业务笔数多而分散、业务流程速度快、期限短、往往无抵质押担保等特征，使得传统的信用评估手段难以奏效。大数据技术的发展，使得信用评级能够广泛采用个人消费者的行为数据构建基于消费者行为的信用评级模型，使信用评估具有响应速度快、结果相对准，从而有可能得到日益广泛的运用。本文即尝试采用大数据方法，以某公司消费金融数据为基础，以机器学习方法构建信用评估模型，并对结果进行评价。

下文首先回顾消费金融与消费金融风险管理的相关文献，进而对样本公司的数据进行分析探讨，并在大数据理念指导下，尝试以两种机器学习算法为工具构建信用评分模型，最后以常见的 LOGISTIC 模型为基准，比对机器学习模型的优劣，最后再给出未来的研究展望。

二．文献回顾

欧美发达国家消费金融业务发达，但与其他金融问题相比，消费金融的研究并不算丰富。Merton(1969)从消费者视角出发将消费金融定义为在给定的金融环境中，利用所掌控的资产配置来最大程度地满足消费者的各种需求，访定义范围最广因此广为接受。Tufano(2009)从满足消费者的金融需求，从使用功能的角度进行研究，将消费金融的范围领域划分为支付消费、信贷消费、防范风险消费、投资性消费，其中信贷与国内消费金融业务范围相似。Allen(2012)指出美国信贷渠道多样化，且家庭信贷规模与家庭消费、授信银行、家庭收入等因素都存在着相关性。Sonenshein(2014)发现信用约束程度与消费者获取贷款难易程度、成本高低都成负相关关系，是信用分层定价的现实体现。Kregel(2016)发现消费金融与国家金融体系与征信体系的发达程度直接相关。

与国外不同，受益于我国互联网经济的快速发展，我国消费金融无论是从实际业务开展上还是学术研究上均得到了飞速发展，国内学者从消费金融的概念、功能、范围、业务模式、面临的风险以及发展趋势等方面对消费金融进行了大量研究。王江(2010)、周美英（2017）、杨才勇、李东耀（2015）、袁天昂（2017）、王雅俊(2017)、曹淼孙（2018）等从消费金融概念、发展历史与趋势、制约因素、监管思路等角度对消费金融分别进行了有益探讨，丰富了

中国消费金融理论。

消费金融信用风险管理作为信用风险管理的一个应用领域，随着信用风险评估、信用风险管理的理论与技术同步发展，陈忠阳（2010）对海外信用风险量化管理理论与技术进行了较为全面的回顾。技术方面，早期研究中经典统计方法在消费金融的风控领域占据主导地位，主要围绕着传统的 Logistic 模型进行分析。于立勇（2004）、张国政（2015）、仵伟强，后其林（2018）等借助不同数据源构建了信用评分的 Logistic 模型，分析了该类型的优劣势。近年来随着大数据技术的发展，采用大数据方法论进行信用风险评估的方法在我国得到了较多探讨与应用。雷晨念（2016）借助 BP 神经网络信用构建风险度量系统，唐剑琴(2016)利用 Lending Club 的数据，构建出了 C4.5 决策树信用风险度量模型；刘厚钦（2019）则创新地引入了用户检测量表，通过用户基本信息、用户检测量表、流水记录等相关数据，借助 GBDT 算法对于用户的违约概率进行预测。顾洁（2017）从消费金融公司的视角，认为互联网消费金融模式对风险控制提出了更高的要求，以 BD 消费金融公司的分期付款业务为研究样本，通过 Logistic 回归算法、决策树算法、Bagging、Boosting、随机森林构建消费分期信贷决策模型，对结果加以比较，发现其余算法的效果较 Logistic 回归算法并未提升，陆健健,江开忠（2019）认为相比随机森林（RF）与 GBDT 算法，建立在 XGBoost 集成算法上的个人信用评估模型性能最优，在准确率指标上明显高出随机森林与 GBDT 算法等。目前，更多的学者关注于现有机器学习算法的组合优化。赵金剑（2017）针对不平衡数据集问题，对 XGBoost 算法进行改进，采用 XGBoost 算法框架与代价敏感相融合的 CS-XGBoost 算法来构建信用逾期风险的预测模型；王重仁，韩冬梅（2017）利用改进的卷积神经网络对互联网金融信用风险进行预测，预测效果明显好于 Logistic 回归与随机森林模型；程玉胜,邹欢（2018）在信用风险预测上，采用随机森林重新构建出新型 RFM(Recency, Frequency, Monetary)模型，并与 C4.5 算法、人工神经网络和 KNN 算法进行了对比；王重仁,路高飞(2019)在对借款者个人信用风险进行研究中，通过遗传算法对传统 BP 神经网络进行优化，发现改进后的神经网络拥有更少的迭代次数，更快的收敛速度，预测准确率也大幅提升。总体而言，消费金融概念在国内兴起较晚，目前更多的机器学习在消费金融领域的研究，采用的多是类消费金融数据或者国外公开数据库的开源数据，完全利用国内消费金融实际业务数据进行建模的研究较少；其次，关于机器学习算法的研究创新性不够，更多的停留在模型的验证对比以及简单的组合优化层面，对模型的应用及其产生的经济效益的分析不足。

三．研究方法

信用评分是评估或预测信用风险的办法，它是根据客户信用历史数据，利用一定信用评分模型得到客户的信用得分或信用级别，其本质上是对客户进行分类。目前有很多方法用于建立信用评分模型，比如 logistic 回归、支持向量机、人工神经网络等。本文拟以 GBDT 算法与 XGBoost 算法建立信用评分模型。

1．GBDT 算法

GBDT（Gradient Boosting Decision）全名为梯度提升决策树模型，是一种高精度的分类器，通过聚集多个基分类器的预测提高分类准确率，常用于解决分类或回归问题。GBDT 采用的方法是提升(Boosting)。假设$f(x)$为是我们的拟合目标，GBDT 算法主要是通过多次迭代寻找到$f(x)$，在迭代过程中，GBDT 通过在残差减少的梯度上建立新的模型来减少残差，算法框架如下：

（1）设损失函数为$\mathrm{L}(\mathrm{y}, f(x))$，如果第 k-1 次迭代得到的估计函数是$f_{k-1}(x)$，那么对应的损失函数则为$\mathrm{L}(\mathrm{y}, f_{k-1}(x))$；

（2）在经过 k 次迭代之后，通过找到弱学习器$g_k(x)$，使得：$\mathrm{L}(\mathrm{y}, f_k(x)) = \mathrm{L}(\mathrm{y}, f_{k-1}(x)) + g_k(x)$取得最小值；

（3）令 N 为最大迭代次数，持续迭代，直到达到迭代终止条件：$\mathrm{L}(\mathrm{y}, \mathrm{f_k(x)})$趋向于 0，最终得到新的强学习器$\mathrm{f(x)} = \mathrm{f_N(x)}$。

2．XGBoost 算法

XGBoost 全名为极端梯度提升，是一种集成学习算法，属于 3 类常用的集成方法(bagging,boosting,stacking)中的 boosting 算法类别。这是一个加法模型，基模型一般选择树模型，但也可以选择其它类型的模型如逻辑回归等。XGBoost 采用二叉树，从开始的全部样本都在一个叶子节点上开始，叶子节点不断通过二分裂，逐渐生成一棵树。XGBoost 使用 level-wise 的生成策略，即每次对同一层级的全部叶子节点尝试进行分裂。XGBoost 采用特征并行的方法进行计算选择要分裂的特征，即用多个线程，尝试把各个特征都作为分裂的特征，找到各个特征的最优分割点，计算根据它们分裂后产生的增益，选择增益最大的那个特征作为分裂的特征。

XGBoost 属于梯度提升树(GBDT)模型这个范畴，GBDT 的基本想法是让新的基模型（GBDT 以 CART 分类回归树为基模型）去拟合前面模型的偏差，从而不断将加法模型的偏差降低。XGBoost 做了一些改进，从而在效果和性能上有明显的提升。

（1）GBDT 将目标函数泰勒展开到一阶，而 XGBoost 将目标函数泰勒展开到了二阶。保留了更多有关目标函数的信息，对提升效果有帮助。

（2）GBDT 是给新的基模型寻找新的拟合标签（前面加法模型的负梯度），而 XGBoost 是给新的基模型寻找新的目标函数（目标函数关于新的基模型的二阶泰勒展开）。

（3）XGBoost 加入了和叶子权重的 L2 正则化项，因而有利于模型获得更低的方差。

（4）XGBoost 增加了自动处理缺失值特征的策略。通过把带缺失值样本分别划分到左子树或者右子树，比较两种方案下目标函数的优劣，从而自动对有缺失值的样本进行划分，无需对缺失特征进行填充预处理。

（5）XGBoost 还支持候选分位点切割，特征并行等，可以提升性能。

四．实证设计

1．数据来源及其处理

（1）数据源

选取 M 信托其中一个助贷机构自 2018 年 10 月 31 日至 2019 年 2 月 20 日期间的用户贷款数据，产品期限为 6 期和 12 期的用户，共计 24655 笔样本数据（脱敏），根据数据表现期完整性以及数据覆盖程度，最终筛选出的 17738 个有效建模样本。利用独立第三方数据源的前期数据积累优势进行联合封闭建模。

（2）数据标签

本次建模结合该助贷机构客户历史表现特征，通常逾期超过 15 天的贷款收回概率较低，逾期超过 30 天的贷款收回概率更低，逾期超过 30 天则可认为是坏账。以此给出好/坏客户定义。

坏客户：最大逾期天数>15 天的客户；

好用户：表现期足够且逾期天数≤15 天的客户。

根据该助贷机构的历史表现，将坏用户定义为逾期超过 15 天相对 30 天更保守，且模型预测更有前瞻性。使用的整体建模样本 17738 笔，其中坏客户 611 个，坏客户浓度 3.44%。

（3）变量选择

通过与第三方数据源特征变量匹配，共 1267 个变量，其中，浮点型（float64）变量共 101 个，整型（int64）变量共 1146 个，对象类型（object）变量共 20 个。剔除 IV（IV，information value，信息价值或信息量，用来衡量变量的预测能力）值小于 0.1 的字段，最终留下 491 个变量（含因变量，即“好”“坏”标签变量）

（4）训练集与测试集的划分

将数据按时间排序并处理后，前 70%作为训练集，后 30%为测试集，以训练集训练评分模型，以测试集验证模型效果。训练集与测试集的“好”“坏样本分布如表 1 所示。

表 1 带“好”、“坏”标签的数据集划分结果

标签描述	训练集	测试集
0	12076	5051
1	341	270

由表 1 可见，训练集和测试集的分布不一致，且坏客户比重随时间而增加。

2．参数设置与优化

本研究以用 python 实现，GBDT 分类算法通过导入 sklearn.ensemble 包的 GradientBoostingClassifier 来实现，XGBoost 分类算法则通过引入 xgboost.sklearn 包的 XGBClassifier 来实现。下面分别说明这两种算法的参数设置与优化方案。

（1）GBDT 算法参数设置与优化

A．学习速率（又称步长）与最大迭代次数。学习速率（步长）和最大迭代次数共同确定算法的拟合效果，因此这两个参数需一并设置与优化，本研究初始值步长 learning_rate=0.01，最大迭代次数设为 100：n_estimators = 100。

B．子采样，默认值为 1.0，即全部数据参与建模，这容易导致过拟合。小于 1 时能减少方差，即预防过拟合，但会导致样本拟合偏差加大，所以也不宜过低。本研究取值 0.7：subsample=0.7。

C．损失函数。对分类模型而言，损失函数取值可为“deviance”（对数似然损失函数）及“exponential”（指数损失函数）两种选择。默认为“deviance”，它对二元分类及多元分类均有较好的优化。本研究取默认值“deviance”：loss = “deviance”。

D．最大特征数。指分类过程中要考虑的最大特征数，以控制决策树生成的时间。默认值为“None”,意为分类划分时考虑所有的特征数；如选取整数值，则考虑特征的绝对数；如选取浮点数值，则考虑特征数的百分比。例如，记样本总量为 N，选取“log2”意为分类划分过程最多考虑$\log_2 N$个特征，选取“sqrt”意为最多考虑$\sqrt{N}$个特征。本研究选择 max_features=“sqrt”。

E．决策树的最大深度。该参数决定决策树子树的深度，如果为空，表示子树的深度将不会受到限制，通常如果样本数据量小或特征数少的时候可以这样设，但如果模型样本数据量大、特征数多时，建议限制决策树最大深度，常用取值范围为[1-100]。本文取 max_depth=1。

F．内部节点再划分时所需最小样本数。此值限制决策树子树的继续划分，若某个节点样本数量少于该值，则不会再继续尝试选择最优特征来进一步划分。该值默认为 2，假如建

模样本量不大，则可忽略此取值，但如果建模样本量的数量级非常大，则建议增大这个值。本文初始取值为 3，min_samples_split=3。

G. 叶子节点最少样本数。若某叶节点数小于样本数，则会连带其兄弟节点一并被剪枝。默认值为 1，假如样本量不大可忽略此取值，但若样本量非常大，则建议增大此值。本研究初始取值为 100，min_samples_leaf=100。

H．叶子节点最小样本权重和。此值对叶子节点所有样本权重和的最小值进行了限制，若小于该值，则会连带其兄弟节点一并被剪枝。通常情况下，若有较多建模样本有缺失值，或分类树样本分布的类别差异较大时，则需引入样本权重，此时需关注调整此值。本研究初始值取默认值 0，min_weight_fraction_leaf = 0。

I．最大叶子节点数。该值通过限制最大叶子节点数来防止模型过拟合。通过调整该值加以限制，则算法会建立最大叶子节点数内的最优决策树。假若特征数不多时，可忽略该取值；但若是特征数分层多的话，则应加以限制，具体取值可根据交叉验证获取。本研究初始对此不加限制，max_leaf_nodes = “None”。

J．节点划分最小不纯度。此值能够控制决策树的增长，若某个节点的不纯度(基于基尼系数，均方差)小于该阈值，则该节点不再继续生成子节点，即成为叶子节点。本研究初始值采用默认值，min_impurity_split = 1e-7。

K．参数优化。Boosting 主要关注降低偏差，即算法的期望预测与真实预测之间的偏差程度，模型调优的过程就在于寻找最优参数组合，从而使得残差不断减小。模型本身具有拟合能力，为使得每次迭代上更加拟合数据，保证测试与训练集之间具有较小的偏差，故选择模型简单，决策树深度选择较小值，这样可防止过拟合。当 KS 残差小于 0.01 时，则可认为调参达到较优结果。

（2）XGBoost 算法参数设置与优化

A．损失函数。该参数定义问题类型及所需的最小化的损失函数。信用评分建模本质是评估户信用好坏的二分类问题，故用此参数用：objective= “binary:logistic”。

B．评测函数。该值用以对不同问题的解决结果度量其优劣，如果是回归问题则用均方根误差 RMSE 或平均绝对误差 MAE，如果是分类问题，则多分类可用多分类错误率 merror 或多分　类损失函数 mlogloss，如果是二分问题，则用二分类错误率 error 或负对数似然函数 logloss，以胶 AUC（即 ROC 曲线下面积）。信用评分问题本质是对客户信用进行二分，因此本研究选择 eval_metric=“auc”。

C．学习速率（步长），learning_rate=0.3。默认值为 0.3，该值越小，则模型训练速度越

慢，取值范围在（0，1]，需根据实际情况调整；

D．最大迭代次数，初始值 n_estimators=80。该值决定决策树的个数，数量越多，越容易过拟合；

E．决策树最大深度，初始值 max_depth=6。默认值为 6。该值可用以防止过拟合。值设置越大，模型就能学到更局部更具体的样本。典型取值范围[3，10]；

F．gamma 值，用以限定节点分裂所需最小损失函数下降值。值越大算法越保守，并且参数值和损失函数息息相关。本研究初始值 gamma=3。

G．最小叶子节点的样本权重和，初始值 min_child_weight=126。默认值为 1。该值决定最小叶子节点样本权重和，用于防止模型过拟合。当值较大时，可以防止模型学习到局部的特殊样本。但若值过高将导致欠拟合；

H．子采样。subsample=0.8。默认值为 1。该值控制每棵树的子采样的比例。若该值减小，则算法变得更加保守，防止过拟合产生。但若值设置得过小，则可能导致模型欠拟合。典型取值范围[0.5,1]；

I．列采样占比，初始值 colsample_bytree=1。该值控制每棵随机采样的列数的占比(每一列是一个特征)，典型取值范围[0.5,1]；

J．权重的 L1 正则化项，初始值 reg_alpha=1.1。。和 Lasso regression 类似，可以应用在很高维度的情况下，使得算法的速度更快；

K．权重的 L2 正则化项，初始值 reg_lambda=28。默认值为 1。。和 Ridge regression 类似，用来控制 XGBoost 的正则化部分；

L．随机数种子，random_state=42。调节此参数，可使模型结果稳定。

M．优化标准，同 GBDT 算法，当 KS 残差小于 0.01 时，则可认为调参达到较优结果。

五．主要结论与建议

1．入模变量及其含义

如附表 1 所示，GBDT 算法选择出的入模变量有 16 个，而如附表 2 所示，通过 XGBoost 算法选择出的入模变量为 27 个。这些变量类型主要集中在以下几个方面。

（1）信用风险识别，整体 IV 值高，其分类为消费分期、现金分期、类信用卡，反映了本次建模客群数据偏向于这三类客群，也反映了这三类客群交叉重合度较高；

（2）反欺诈，体现在信用卡代偿类的反欺诈风险识别，说明建模样本中有恶意欺诈用

户存在；

（3）借贷意向（多头借贷），此类数据变量入模个数最多。GBDT 模型中 16 个变量里占 9 个，XGBoost 模型中占 27 个变量中的 11 个，这反映了多头风险严重且不断累积，越来越成为消费金融风险的核心风险。

机器学习算法通常是“黑盒”训练，选出的变量较多，且解释性不强。但本案例中 GBDT 算法根据叶子节点分裂时的不纯度迭代来筛选变量，给各特征变量赋予不同的重要性权重，权重高则特征影响程度大，而 XGBoost 算法本质是基于决策树的集成算法，根据特征重要性（Feature Importance）来筛选变量，按照 F_score 特征重要性评分进行排序，得分高则特征影响程度大，因此两类模型的大多数变量均具有较好的可解释性。而 XGBoost 算法入模变量还有第四个方面，既客户行为相关变量，如客户浏览等级、两年家具建材类浏览次数、近两年服装配饰类消费总金额。这类数客户行为数据的解释性稍弱，但其背后一定也存在其经济学意义，有待后续挖掘。

2．模型区分效果

GBDT 建模结果如图 1 图 2 所示，K-S 接近 0.39，AUC 达到 0.77，XGBoost 建模的结果如图 3 和图 4 所示，K-S 达到 0.42，AUC 达到 0.78，两类模型对于好坏用户都有很好的区分度，并且训练集和测试集效果相当，区分效果都相当稳定。

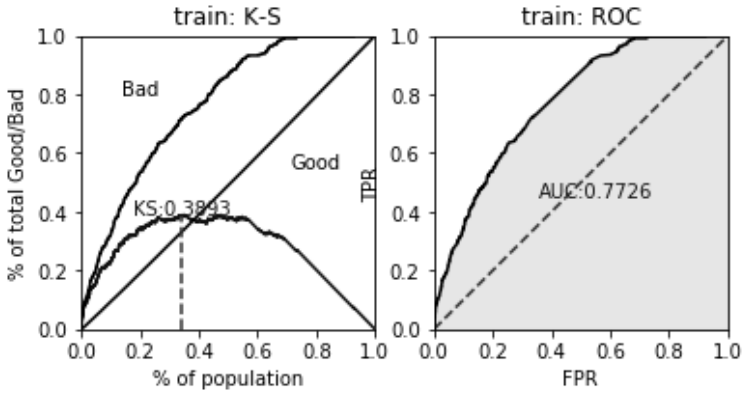

图.1 GBDT 建模训练样本的 K-S 图及 ROC 图

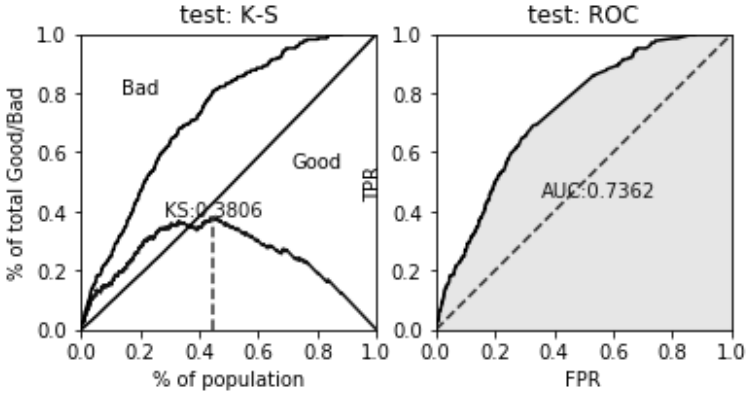

图 2 GBDT 建模测试样本的 K-S 图及 ROC 图

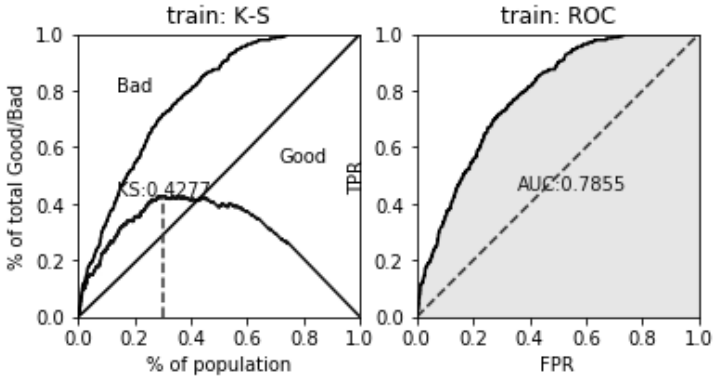

图 3 XGBoost 建模训练样本的 K-S 图及 ROC 图

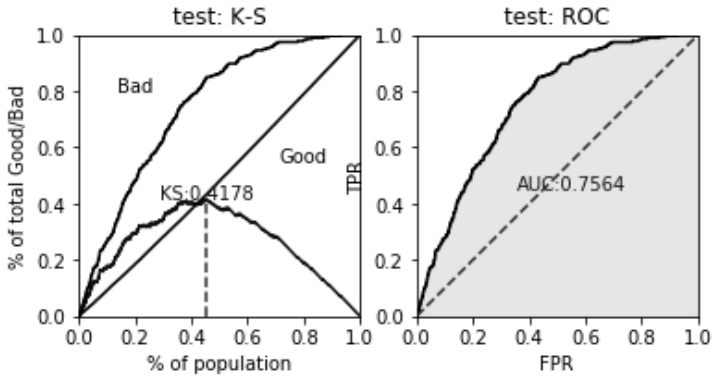

图 4 XGBoost 建模测试样本的 K-S 图及 ROC 图

为验证 GBDT 算法及 XGBoot 算法的区分效果，本研究构建了 Logistic 模型，其区分度及模型稳健性如表 2 所示。

表 2　Logistic 回归、GBDT、XGBoost 模型结果对比

算法	K-S-训练集	K-S-测试集	AUC-训练集	AUC-训练集	PSI
Logistic 回归	0.3755	0.3633	0.7486	0.7396	0.0041

GBDT	0.3893	0.3806	0.7726	0.7362	0.0028
XGBoost	0.4277	0.4178	0.7855	0.7564	0.0046

由表 2 可见，三种模型的训练集以及测试集的 K-S 值均大于 0.3，说明区分度较好；训练集以及测试集的 AUC 值均大于 0.7，可以判断预测分类器的分类效果较好；稳定性指数 PSI 均小于 0.01，说明三个模型稳定性都很高；三个模型中，GBDT 算法模型的 K-S 值、AUC 值均优于 Logistic 回归模型，在效果表现上较好，而 XGBoost 测试集的 AUC 值为 0.7564，表现最优，总体而言，GBDT 算法和 XGBoost 算法结果优于 logistic 模型。

3．模型应用的经济效益

应用三种模型对样本进行回溯，各模型选择其最优阈值进行判别，客户拒绝率 3与逾期坏账率 4的模型对比结果如表 3 所示，GBDT 模型与 XGBoost 模型在拒绝率更少的条件下，逾期坏账率更低，更直观地反映了这两个模型预测能力更强。

表 5.1 评分模型的拒绝率与逾期坏账率回溯结果对比

预测概率 / 算法模型	拒绝率	逾期坏账率
Logistic 回归模型	19.81%	1.88%
GBDT 模型	19.55%	1.87%
XGBoost 模型	18.67%	1.84%

4．建议

上世纪 80 年代,美国费埃哲公司以 logistic 回归算法构建了费埃哲信用评分体系,成为美国信用评分市场领头羊，Logistic 回归在我国信用评分市场也得到广泛运用。随着大数据技术的迅猛发展，新算法、新技术的层出不穷，以大数据技术为基础的信用评分模型探索也日益繁多。本研究仅仅是以某信托机构消费金融数据为基础进行了一次评分建模尝试，结果表明，大数据模型确实有优于传统 Logistic 模型。但是，大数据不是一个神奇的机器，它只是模型构建的工具，其结果并不绝对，根据数据特征和算法特性构建合适的模型非常关键，数据科学团队对算法核心原理的深刻理解、快速的算法实现能力、强大的大规模数据处理能力，对于充分利用大数据算法开发出高性能的信用风险评估模型非常重要。而在实际建模开发与布署的过程，业务专家与数据科学团队在数据逻辑的理解和建模指标的选取上的紧密合作更是绝对重要的一环。

[3] 拒绝率：拒绝申请笔数占总贷款申请笔数的比例。

[4] 逾期坏账率：仍然以逾期 15 天以上客户的逾期笔数占总贷款申请笔数的比例。

附表一：GBDT 模型入模变量

变量英文名	变量中文名	iv	重要性权重
scoreconson	信用风险识别-线上消费分期	0.894271	0.210881
scorecashon	信用风险识别-线上现金分期	0.778758	0.055613
scorerevoloan	信用风险识别-信用卡（类信用卡）	0.733899	0.088799
scoreafcreditbt	反欺诈风险识别-信用卡（类信用卡）-信用卡代偿	0.522178	0.269344
alf_apirisk_d30_sum	最近 30 天申请的申请机构风险等级求和	0.448909	0.01737
ir_m12_id_x_cell_cnt	近 12 月身份证关联手机号个数	0.438645	0.117288
alf_apirisk_d15_sum	最近 15 天申请的申请机构风险等级求和	0.386364	0.016589
als_max_m12_id_cell_nbank_p2p_allnum	近 12 个月在非银机构-p2p 机构申请次数	0.24767	0.011869
ir_m1_id_x_device_cnt	近 1 月身份证关联设备个数	0.22095	0.071006
alm_d7_id_nbank_orgnum	按身份证号查询，近 7 天内在非银机构申请机构数	0.217187	0.011269
alm_d7_cell_nbank_allnum	按手机号查询，近 7 天内在非银机构申请次数	0.21174	0.01962
ir_m3_id_x_device_cnt	近 3 月身份证关联设备个数	0.190833	0.027419
alm_d15_cell_nbank_ca_orgnum	按手机号查询，近 15 天内在非银机构-现金类分期申请机构数	0.166724	0.03021
alm_m1_cell_nbank_ca_orgnum	按手机号查询，第 1 个月内在非银机构-现金类分期申请机构数	0.164483	0.00857
alm_m7_id_nbank_orgnum	按身份证号查询，第 7 个月内在非银机构申请机构数	0.146912	0.009919
als_max_m3_id_cell_nbank_oth_orgnum	近 3 个月在非银机构-其他申请机构数	0.139677	0.034235

附表二：XGBoost 模型入模变量

变量名英文名	变量中文名	IV 值	F_score
scorecashon	信用风险识别-线上现金分期	0.78	4
scorerevoloan	信用风险识别-信用卡（类信用卡）	0.73	4
scoreconson	信用风险识别-线上消费分期	0.89	3
ir_m12_id_x_cell_cnt	近 12 月身份证关联手机号个数	0.44	3
pc_noregincome_lst_mons	无稳定工作时长	0.14	2
alf_apirisk_all_sum	过去全部申请的申请机构风险等级求和	0.91	2
als_max_m3_id_cell_nbank_else_orgnum	近 3 个月在非银机构-其他申请机构数	0.21	2
alf_apirisk_d360_mean	最近 360 天申请的申请机构风险等级均值	1.57	2
pc_business_type	行业类别预测	1.05	2
alf_time_intedays_d30_mean	最近 30 天申请时间间隔均值	0.99	2
alf_time_intedays_d360_mean	最近 360 天申请时间间隔均值	1.55	2
scoreafrevoloan	反欺诈风险识别-信用卡（类信用卡）	0.34	2
alf_apirisk_d180_mean	最近 180 天申请的申请机构风险等级均值	1.47	2
scoreafcreditbt	反欺诈风险识别-信用卡（类信用卡）-信用卡代偿	0.52	2
als_max_m12_id_cell_nbank_p2p_allnum	近 12 个月在非银机构-p2p 机构申请次数	0.25	2
als_max_m6_id_cell_nbank_cons_allnum	近 6 个月在非银机构-持牌消费金融机构申请次数	0.28	2
scoreafconsoff	反欺诈风险识别-线下消费分期	0.2	1

scoreconsoff	信用风险识别-线下消费分期	0.62	1
cf_cons_C9_views	近两年家具建材类浏览总次数	1.2	1
cf_cons_C7_amount	近两年服装配饰类消费总金额	0.45	1
scorecreditbt	信用风险识别-信用卡（类信用卡）-信用卡代偿	0.85	1
cf_prob_mean	请求 id 与 cells 的平均置信度	1.01	1
alf_apirisk_d360_sum	最近 360 天申请的申请机构风险等级求和	0.83	1
cons_tot_m12_visits	近 12 个月浏览等级	0.79	1
tl_m6_cell_nobank_allorgnum	按手机号查询，第 6 个月内在非银机构借贷机构数	0.11	1
frg_list_level	查询人欺诈团伙等级	0.21	1
tl_m1_cell_nobank_passnum	按手机号查询，第 1 个月内在非银机构新增核准借贷次数	0.33	1

中国人民大学国际货币研究所
INTERNATIONAL MONETARY INSTITUTE OF RUC

地址：北京市海淀区中关村大街 59 号文化大厦 605 室，100872　电话：010-62516755　邮箱：imi@ruc.edu.cn