计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(10):382−388 [doi: 10.15888/j.cnki.csa.008724]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

信用逾期预测中不同机器学习模型对比分析①

陈　霞

(中国人民大学 统计学院, 北京 100872)
通信作者: 陈　霞, E-mail: sara_chenx@126.com

摘　要: 当前金融机构正在努力应对不良资产的增长问题, 在信贷领域借贷逾期预测结果的准确性将直接决定不良资产的规模. 为了更好预测借贷人的还款能力, 通常会引入数据模型方法, 但对于数据样本较少的新业务, 单纯用这类数据容易导致模型结果过拟合. 本文通过实际案例分析, 对小样本业务数据进行相似业务数据补充, 并采用随机森林、LightGBM、XGBoost、DNN 和 TrAdaBoost 迁移学习方法, 旨在为小样本业务在模型建立过程中样本不足的问题提供一种有效的解决方法. 研究结果表明, 针对数据量少的产品, 结合相似金融业务数据后采用这五种机器学习模型方法的预测结果 AUC (area under curve) 均大于 80, 其中使用迁移学习模型比 LightGBM、XGBoost、DNN 和随机森林模型在预测集上的 AUC 至少高出 2 个点; 此外迁移学习模型的预测结果的精准率 (88%) 和召回率 (73%) 也是最高的.

关键词: 小样本; 信贷业务; 逾期风险; 机器学习模型; 风险预测

引用格式: 陈霞.信用逾期预测中不同机器学习模型对比分析.计算机系统应用,2022,31(10):382–388. http://www.c-s-a.org.cn/1003-3254/8724.html

Comparison Analysis of Different Machine Learning Models in Credit Overdue Prediction

CHEN Xia

(School of Statistics, Renmin University of China, Beijing 100872, China)

Abstract: Financial institutions are currently grappling with the growth of non-performing assets (NPAs). The prediction accuracy of credit overdue directly determines the size of NPAs. For better prediction of repayment ability, data modeling methods are often introduced, which may cause over-fitting for new businesses with small data samples. This study performs case studies and enriches the small data samples by similarity with random forest, LightGBM, XGBoost, DNN, and TrAdaBoost transfer learning. It aims to provide an effective solution to insufficient samples during the model establishment for small sample businesses. The results show that the area under curve (AUC) of the five machine learning models is greater than 80 for small data samples after similar financial business data are integrated. The AUC of TrAdaBoost is at least 2 points higher than that of LightGBM, XGBoost, DNN, and random forest models on the prediction set. In addition, TrAdaBoost stands out with the highest precision (88%) and recall (73%).

Key words: small sample; credit business; delinquency risk; machine learning models; risk prediction

风险控制是衡量金融行业是否健康可持续发展的重要因素, 一直也是金融公司重点研究的内容. 当前金融信贷业务量随着消费升级不断高涨, 但是违约风险也在日趋凸显, 如某些小型贷款机构不得不依赖自己在银行的担保金勉强维持. 据公开数据研究, 中国上市的商业银行不良贷款余额逐年增长, 并在 2020 年达到了历史最高. 四大国有行合计坏账万亿元, 居商业银行首位, 占上市银行不良贷款总额约 6 成. 截至 2020 年

① 收稿时间: 2022-01-03; 修改时间: 2022-01-29; 采用时间: 2022-02-22; csa 在线出版时间: 2022-06-28

末, 中国工商银行不良贷款余额排名第一, 其次为中国建设银行、中国农业银行和中国银行, 不良贷款余额均高于 2 000 亿元[1]. 这些结果充分反映了银行由于没有控制好风险导致了巨额不良资产的问题, 因此金融公司在开展贷款业务时应将风险控制放在首要位置.

为了更好地控制业务风险使自身获利, 金融机构不断挖掘存量用户特征信息, 以此区分好用户和坏用户. 早在 20 世纪 90 年代开始, 金融公司为了获利, 把各类统计分析算法应用在业务中, 通过模型拟合的方法提前判断出用户风险[2–5]. 但对于数据样本较少的新业务, 单纯用这类数据容易导致模型结果过拟合. 本文试图结合相似金融业务数据做为模型训练样本, 运用目前金融行业运用较多的算法: 随机森林、LightGBM、XGBoost、DNN 和迁移学习, 分别预测新业务出的结果并与真实结果进行比较, 旨在为小样本业务在模型建立过程中样本不足的问题提供一种有效的解决方法.

1 信贷业务在模型上的发展情况

得益于 Nasdap 系统, 1971 年美国的互联网金融进入正式运营, 1995 年美国成立了一家网络银行, 从此互联网金融进入了发展期. 20 世纪 90 年代开始, 发达国家在互联网金融领域快速发展, 互联网金融服务逐渐多元化、综合化, 行业之间竞争非常激烈. 各公司为了提升利润, 降低风险逾期率迫在眉睫, 各种统计分析算法应用在金融风控中, 大数据量化风控成为主流思想, 如在信贷引入决策树模型、逻辑回归模型、判别分析以及 BP 神经网络模型[2–5] . 由于逻辑回归模型可解释性较强, 在金融领域备受青睐, 然而逻辑回归算法要求数据满足严格的假设, 因此在实践中很难应用[6]. 相比于逻辑回归模型, 随机森林、LightGBM 和 XGBoost 等树模型采用集成模型的思想, 拟合效果更好. DNN 深度学习模型则可在稀疏空间做分类, 通过增加节点数或激活函数的次数来增加线性或者非线性转换能力和次数, 且尽可能的优化损失函数去学习规则, 但其解释性相对较差.

为了满足信贷模型预测效果更好的要求, 可从模型算法、数据输入和变量挖掘 3 个方面来进行优化. 模型方面可以优化模型算法或是利用组合模型进行预测, 如使用不同核函数建立支持向量机模型、基于 XGBoost 机器学习算法建模、使用加权投票法建立组合模型、基于梯度提升决策树模型、建立 SVM-Logistic 组合模型、建立随机森林等与逻辑回归融合模型[7–12]; 变量方面可扩大量化维度, 如蒋翠清等[13] 将借款用途和社交情况等信息进行量化, 分析了不同软信息对贷款违约的影响作用; 数据方面可进行抽样等操作, 如祝钧桃等[14] 针对小样本数据从数据增强、度量学习、外部记忆、参数优化 4 个方面解决小样本问题, 为往后的研究提供了有价值的参考.

2 预测模型方法和数据来源

该实战案例分析使用 iOS 系统 10.14 版本, 软件为 Jupyter notebook; 具体硬件配置: 内存 8 GB、处理器为 2.3 GHz Intel Core i5; 实验中使用的工具为 Python 3.7 Sklearn、TensorFlow、Kears 等.

2.1 基础数据来源

数据源来自某银行信贷业务, 分为历史金融贷款数据和现业务数据, 历史信贷数据时间范围为 2014 年 1 月–2017 年 12 月, 按天记录, 共 30 万条数据. 当前金融业务数据共 1.5 万条, 时间范围为 2017 年 1 月–2017 年 12 月. 由于需要大规模开展业务, 需要结合历史信贷数据评估业务风险, 如通过用户的历史逾期情况、资产负债比例、工作年限等维度, 用于预测个人信誉问题.

该实验数据离散变量主要包括有工作年限、工作行业和房产情况. 连续变量数据情况如表 1 所示.

2.2 预测模型方法

(1) 随机森林算法

随机森林 (random forest, RF) 模型是 2001 年由 Breiman[15] 提出的基于分类树的算法.它通过对大量分类树的汇总提高了模型的预测精度, 是取代神经网络等传统机器学习方法的新的模型, 在医学、气象、金融、水利等领域被广泛使用.

在算法上, 随机森林是采用 bootstrap sample 方法, 有放回的抽样方式进行数据选择, 然后从所有属性中随机选择 m 个属性. 采用树模型训练模型但没有剪枝过程, 每棵树都尽最大程度成长. 重复 k 次, 建立 k 个模型, k 个模型形成决策森林, 每棵树都是一个弱分类器, 最终的预测结果采用投票的方式整合 k 个弱分类器结果完成预测. 从整体来看, 单棵树存在过拟合、准确度不高、不稳定的现象, 多棵树共同决策可提升模型稳定性和精度. 算法步骤如下:

输入为样本集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$ 迭

代 100 次, 输出最终的强分类器$f(x), t=1,2,3,\cdots,100$; 对训练集进行 100 次随机采样, 共使用有放回采集 24.8 万次, 得到包含 24.8 万个样本的采样集合D_t; 用采样集D_t训练第 t 个弱学习器$G_t(x)$, 选择一个最优的特征值作为左右决策的划分点; 分类算法, 则 100 个弱学习器进行投票; 回归算法, 则 100 个弱学习器通过算法平均的方法, 最终拟合出模型结果. 本文采用分类算法完成投票.

随机森林算法参数配置: 100 个弱学习器, 有放回抽样 bootstrap=true, criterion="gini".

表 1 连续变量数据情况

变量名	中文含义	平均值	标准差
debt_loan_ratio	债务收入比	17.53	14.22
del_in_18month	借款人过去18个月逾期30天以上的违约事件数	0.31	0.87
early_return	借款人提前还款次数	1.29	1.45
early_return_amount_3mon	近3个月内提前还款金额	335.23	635.11
f1	金融相关变量f1	0.00	0.04
f2	金融相关变量f2	8.47	7.32
f3	金融相关变量f3	14.66	8.26
f4	金融相关变量f4	8.10	4.87
interest	当前贷款利率	13.22	4.88

注: 数据清理主要遵循以下原则: 缺失值的处理: 缺失值作为一个类别、数据缺失大于50%的变量剔除; 通过psi稳定性判断筛选变量, 稳定性psi大于0.01的变量剔除; 剔除涉及产品本身的变量如期限、贷款金额、贷款时间等.

(2)XGBoost 算法

XGBoost 是基于 GBDT 算法的提升, GBDT 算法仅支持 CART 基分类器, XGBoost 支持 CART 基分类器的基础上同时支持线性分类器. 在精度提升方面, XGBoost 使用二阶泰勒展开式$f(x)=f(x_0)+f'(x_0)(x-x_0)+\frac{f''(x_0)}{2!}\times(x-x_0)^2$, 比 GBDT 更好的逼近损失函数 (loss function). 为了防止过拟合, XGBoost 算法一方面代价函数里加入了正则项来控制模型复杂度, 另一方面借鉴了随机森林的做法, 支持列抽样. 具体算法如算法 1.

算法 1. XGBoost 算法

输入: I, instance set of current node; d, feature dimension
$gain\leftarrow 0$
$G\leftarrow\sum_{i\in I}g_i, H\leftarrow\sum_{i\in I}h_i$
For k=1 to m do
  For j in sorted(I, by x_{jk}) do
  $G_L\leftarrow 0, H_L\leftarrow 0$
  $G_L\leftarrow G_L+g_i, H_L\leftarrow H_L+h_j$
  $G_R\leftarrow G-G_L, H_R\leftarrow H-H_L$
  $score\leftarrow\max\left(score,\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{G^2}{H+\lambda}\right)$
  end
end
输出: Split with max score

XGBoost 算法参数配置: 采用二元分类逻辑回归的方法, 训练 100 次, 最大树深度为 3, 学习率为 0.01, 正则化权重 L1 和 L2 为 1.

(3) LightGBM 算法

LightGBM 在 XGBoost 的基础上做了改进, 主要引入了 Histogram 算法, 内存消耗低并且可快速寻找树的分裂节点. LightGBM 结合单边梯度采样 (gradient-based one-side sampling) 和互斥特征合并 (exclusive feature bundling), 在减少维度和下采样上面进行优化使 Histogram 算法效果更好. 在树的生长上, LightGBM 抛弃了 Level-wise 策略采用 leaf-wise, 为防止过拟合, 使用最大树深限制, 如算法 2 所示.

算法 2. LightGBM 算法

输入: I: training data, d: iterations; a: sampling ratio of large gradient data; b: sampling ratio of small gradient data; $loss$: loss function, L: weak learner
$models\leftarrow\{\}, fact\leftarrow\frac{1-a}{b}$
$topN\leftarrow a\times len(I), randN\leftarrow a\times len(I)$
For i=1 to do
  $preds\leftarrow$ mod $els.predict(I)$
  $g\leftarrow loss(I, preds), w\leftarrow\{1,1,\cdots\}$
  $sorted\leftarrow GetSortedIndices(abs(g))$
  $topSet\leftarrow sorted[1:topN]$
  $randSet\leftarrow RanddomPick(sorted[topN:len(I)], randN)$
  $usedSet\leftarrow topSet+randSet$
  $w[randSet]\times= fact$ Assign weight to the small gradient data.
  $newModels\leftarrow L(I[usedSet], -g[usedSet], w[usedSet])$
  $Models.append(newModel)$

大量的金融信贷场景研究案例表明, LightGBM 在预测结果上表现的效果优于 XGBoost、Logistic、SVM 和随机森林等模型效果, 准确性较高的同时具有较好的鲁棒性[16,17].

LightGBM 算法训练参数设置: 采用 GBDT 提升算法类型, 弱学习器数量为 100, 最大树深度为 3, 学习率为 0.01, 正则化权重 L1 和 L2 为 1.

(4) TrAdaBoost 迁移学习算法

迁移学习将某个领域或任务上学习到的知识或模

式应用到不同但相关的领域或问题中, 适用于源数据较少的场景或者零数据冷启动场景. 迁移学习分为归纳式迁移、直推式迁移、无监督迁移 3 种. 依据迁移知识的形式可将迁移学习分为基于实例的迁移学习、基于特征的迁移学习、基于模型的迁移学习、基于关系的迁移学习.

在 21 世纪初, Ben-David Schuller[18] 提出了学习与任务之间具有相互联系的观点, 为迁移学习提供了理论基础. 利用迁移学习思想在医学上取得了显著成就, 如基于 X 射线和 CT 图像预训练的 CNN 模型进行 COVID-19 检测任务; 把基于自然图像预训练得到的不同 ResNet 模型迁移到乳腺癌诊断任务; 使用与目标数据相似的脑血管图像在 AlexNet 上进行预训练, 再利用 SVM 分类器进行微调训练[19–21]. 在文本挖掘上也常常采用迁移学习方法, 如采用迁移学习方法实现交叉语言文本分类; 利用完善的英文标签处理中文标签缺失问题, 解决了交叉语言迁移分类问题[22,23]. 迁移学习方法在 P2P 信贷实验上表明迁移学习模型的平均 AUC 比逻辑回归模型高 0.088 0, 比支持向量机模型高 0.035 5[24].

TrAdaBoost 迁移学习算法[25] 利用对 AdaBoost 算法加以改来达到迁移学习的效果, 主要通过 boosting 的作用建立自动调整权重的机制, 加重正确的辅助数据权重, 减少不重要的辅助训练数据权重. 主要方法如下:

输入: 两个数据集T_a和T_b, 合并的训练数据集$T=T_a\cup T_b$, 基本分类算法 Learner 和迭代次数 N.

初始化:

1. 初始权重向量$w^1=(w_1^1,\cdots,w_{n+m}^1)^a$, 其中,

$$w_i^1=\begin{cases}\frac{1}{n}, i=1,\cdots,n\\ \frac{1}{m}, i=n+1,\cdots,n+m\end{cases}$$

2. 设置$\beta=1/(1+\sqrt{2\ln n/N})$

For $t=1,\cdots,N$

设置P^t满足$P^t=w^t/\sum_{i=1}^{m+n}w_i^t$, 调用 Learner, 根据合并后的训练数据 T 以及 T 上的权重分布P^t和未标注数据 S, 得到一个 S 的分类器 $h_t:X\mapsto Y$. 计算h_t在T_b上的容错率: $\delta_t=\sum_{i=n+1}^{n+m}\frac{w_i^t|h_t(x_i)-c(x_i)|}{\sum_{i=1}^{n+m}w_i^t}$

设置$\beta_t=\delta_t/(1-\delta_t)^b$

设置新的权重向量:

$$w_i^{t+1}=\begin{cases}w_i^t\beta^{|h_t(x_i)-c(x_i)|}, i=1,\cdots,n\\ w_i^t\beta_t^{-|h_t(x_i)-c(x_i)|}, i=n+1,\cdots,n+m\end{cases}$$

输出最终分类器:

$$h_f(x)=\begin{cases}1, \sum_{t=[N/2]}^{N}\ln\left(\frac{1}{\beta_t}\right)h_t(x)\geqslant 1/2\sum_{t=[N/2]}^{N}\ln\left(\frac{1}{\beta_t}\right)\\ 0, \text{other}\end{cases}$$

TrAdaBoost 迁移学习算法训练参数设置: 基本分类算法采用 XGBoost 模型算法, 并用二元分类逻辑回归训练, 迭代次数为 100, 最大树深度为 3, 学习率为 0.01, 正则化权重 L1 和 L2 为 1, TrAdaBoost 权重修改次数为 8 次, 即训练整体训练次数为 8 次.

(5) DNN 算法

DNN (deep neural network) 神经网络模型又叫全连接神经网络是基本的深度学习框架, 最早由 Hinton 等人[26] 于 2006 年提出, 可基于对数据进行表征学习, 同时能够学习出高阶非线性特征, 具有特征交叉能力. 神经网络总体可分为 3 个模块: 输入层、隐藏层和输出层. 目前应用场景较为广泛, 如图像识别、声音识别、广告推荐、风险预测和智能投顾等场景[27,28]. 本文 DNN 模型结构如图 1 所示.

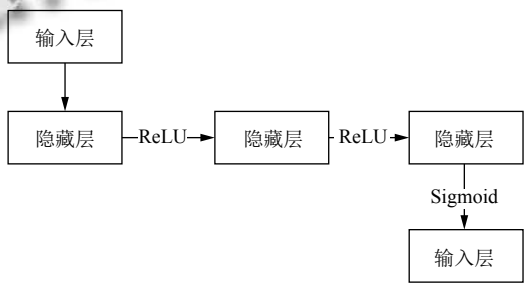

图 1 DNN 模型结构图

具体逻辑如下:

对客户的数据进行预处理, 包括数据清洗和数据变换. 通过 3 层隐藏层后输出预测结果. 每一层可以有一个或多个神经元, 文中模型隐层神经元选用 8 个, 输出层只有 1 个神经元. 激活函数包括 tanh、ELU (exponential linear units)、Sigmoid、ReLU 和 maxout 等, 本文选择 ReLU 函数. ReLU 函数能克服梯度消失的问题, 使得神经网训练速度更快. 输出层设置了 1 个神经元, 使用 Sigmoid 作为激活函数, 输出在 0 和 1 之间.

$$\sigma(x)=1/(1+\mathrm{e}^{-x})$$

由于本文针对金融信贷逾期, 可抽象为好坏预测的二分类问题, 故采用 binary cross_entropy 作为损失函数.

$$loss=-\sum_{i=1}^{n}\widehat{y_i}\log(y_i)+(1-\widehat{y_i})\log(1-\widehat{y_i})$$

其中, $\widehat{y_i}$表示预测结果, y_i表示真实值, 当$\widehat{y_i}=y_i$时, $loss$=0, 表示没有损失.

DNN 算法训练参数设置: SGD 学习率为 0.1, SWA 采用周期性学习, 学习长度 c 为 20, 学习率α_1为 0.001. 在训练过程中, 模型初始化参数之后使用 SGD 进行梯度下降, 迭代 20 个 epoch 后, 将模型的参数进

行加权平均后得到组合权重的集成模型.

2.3 模型评价指标

TP 与 *TN* 表示都分对的情况, *TP* 是样本为正, 预测结果为正; 样本为负, 预测结果为负; *FP* 表示样本为负, 预测结果为正; *FN* 表示样本为正, 预测结果为负. AUC (area under curve) 为 ROC 曲线下与坐标轴围成的面积, AUC 越接近 1.0, 检测方法真实性越高; 当 AUC=0.5 时, 则真实性最低, 则无应用价值.

ROC 曲线的横坐标表示伪正类率, 表示预测为正但实际为负的样本占所有负例样本的比例; 伪正类率即为 *FPR* (false positive rate).

$$FPR = \frac{FP}{(FP+TN)}$$

ROC 曲线的纵坐标为真正类率, 表示预测为正且实际为正的样本占所有正例样本的比例. 真正类率即为 *TPR* (true positive rate).

$$TPR = \frac{TP}{(TP+FN)}$$

精准率 (*accuracy*) 表示正确预测为正和正确预测为负的结果数量占所有预测结果数量的比例.

$$accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)}$$

召回率 (*recall*) 表示正确预测为负的数量占全部负样本数量的比例.

$$recall = \frac{TP}{(TP+FN)}$$

图 2 为本文流程图.

3 模型预测结果对比分析

考虑到需要预测的金融业务数据共 1.5 万条, 则其中 1 万条数据用于模型训练, 5 000 条数据用于模型预测. 目标业务数据样本较少, 结合历史相似信贷模型的 30 万条数据, 模型训练样本共 31 万, 跨时间预测数据共 5 000 条. 坏样本选择逻辑为自放贷后 12 个月的表现期中, 逾期 90 天及以上的用户. 建模数据好坏样本分布情况如表 2 所示.

随机森林算法、XGBoost 算法、LightGBM 算法和 DNN 算法在数据训练时采用 80% 训练, 20% 预测的方法, 为防止模型过拟合, 树模型深度最大为 3. TrAdaBoost 算法中训练集为 30 万历史信贷数据, 预测集目标信贷业务 1 万条数据. 最终模型评价测试数据均为小业务数据, 共 5 000 条. 建模数据测试训练测试数据分布如表 3 所示.

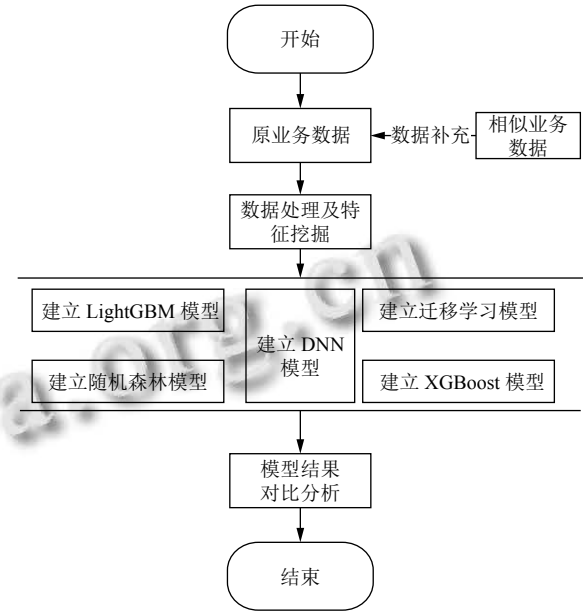

图 2 流程图

表 2 建模数据好坏样本分布情况

数据源	总量	好样本数量	坏样本数量	坏账率 (%)
历史信贷数据	300 000	240 028	59 972	20
目标业务训练集	10 000	8 317	1 683	17
目标业务测试集	5 000	3 418	1 582	32

表 3 建模数据训练测试数据分布

算法	模型拟合数据	测试数据
随机森林	310 000 (80%训练, 20%预测)	5 000
XGBoost	310 000 (80%训练, 20%预测)	5 000
LightGBM	310 000 (80%训练, 20%预测)	5 000
DNN	310 000 (80%训练, 20%预测)	5 000
TrAdaBoost	310 000 (30万训练, 1万测试)	5 000

随机森林、XGBoost、LightGBM、DNN 和 TrAdaBoost 算法预测数据 ROC 曲线结果如图 3 所示, 5 种模型 AUC 结果分别为 84、81、83、84 和 86. 其中 TrAdaBoost 算法效果最好, AUC 的预测结果为 86, 比随机森林和 DNN 的 AUC 高 2 个点, 比 XGBoost 的结果高 5 个点.

表 4 说明了各种算法预测结果的准确率及召回率, 从模型的准确率和召回率来看, TrAdaBoost 算法准确率能达到 88%, 召回率 73%, 均比其他模型效果好; 其次是 DNN, 准确率为 86%, 召回率为 70%; 随机森林算法, 准确率为 84%, 召回率为 68%; 相比于随机森林算

法, XGBoost 算法和 LightGBM 算法对预测数据的召回率更好, 分别是 70%、71%, 其中 LightGBM 算法的准确率比 XGBoost 算法高 1 个百分点.

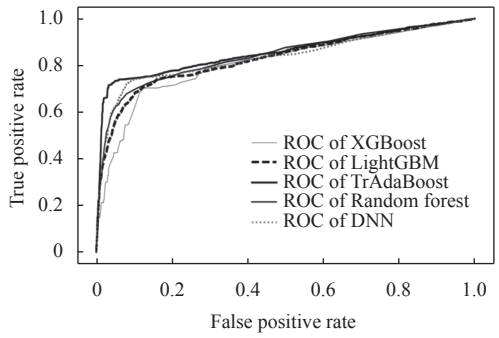

图 3 各算法 ROC 曲线对比图

表 4 各算法预测结果对比 (%)

模型算法	准确率	召回率
随机森林算法	84	68
XGBoost算法	81	70
LightGBM算法	82	71
DNN算法	86	70
TrAdaBoost算法	88	73

图 4 的 TrAdaBoost 算法模型结果分布表明, 模型效果较显著. 把坏账户进行分数的转换后, 按照等量划分的方法把结果分为 8 份, 每份约 1 250 条数据, 黑色的曲线表示坏账率, 可以看出坏账率有下降的趋势, 尤其是前两个区间的坏账率尤其高, 在业务中可以按照这个阈值作为 cut 节点来为业务作辅助决策. 从入模变量的重要性来看, 重要性变量集中在金融属性较强的变量上, 比如借款人提前还款次数和近 3 个月内提前还款金额, 从这两个变量从一定程度上可以说明借款人的财务状况.

4 结论

本研究的主要目的是在银行新开金融产品数据集很小的情况下, 开发一个能对用户是否逾期作出预测的有效模型. 对于金融机构想预判用户是否有逾期风险, 但由于资源的限制, 阻碍了他们获得有效用户数据的管理者来说具有非凡的意义. 把小样本融于其他类似的金融数据集中, 提高模型的预测能力, 对新金融业务具有很强的数据参考价值. 本文研究结果表明, 小样本业务结合相似业务构建模型的思路是可行的. 随机森林、XGBoost、LightGBM、DNN 和 TrAdaBoost 五种算法在测试集上 AUC 结果都高于 80, 精准度也都高于 80%, 召回率平均能达到 70% 以上, 其中 TrAdaBoost 算法 AUC 结果为 86, 精准率为 88% 的情况下召回可达 73%, 效果最好. 总体而言, TrAdaBoost 算法相较于其他对比方法鲁棒性较好, 在预测集上的结果表现最佳. 但是, 本研究在数据的选择上仍有一些缺陷, 例如, 在入模变量的数据选择上只用了银行内部的数据, 未引入三方数据而导致用户画像不全, 使得预测集的准确率和召回率还有提升空间, 后面可进一步补充民间借贷等相关数据.

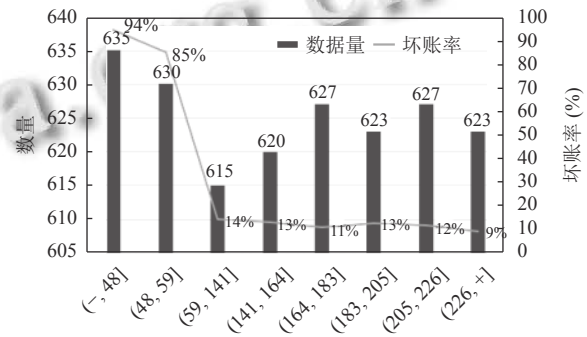

图 4 TrAdaBoost 算法模型结果区间和坏账率分布图

参考文献

1 王立峰. 2021 商业银行坏账报告: 1.7 万亿不良, 工农中建金额居首. 红周刊, 2021.

2 Hand DJ, Henley WE. Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 1997, 160(3): 523–541. [doi: 10.1111/j.1467-985X.1997.00078.x]

3 Thomas LC. A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. International Journal of Forecasting, 2000, 16(2): 149–172. [doi: 10.1016/S0169-2070(00)00034-0]

4 Altman EI. The importance and subtlety of credit rating migration. Journal of Banking & Finance, 1998, 22(10–11): 1231–1247.

5 Tam KY. Neural network models and the prediction of bank bankruptcy. Omega, 1991, 19(5): 429–445. [doi: 10.1016/0305-0483(91)90060-7]

6 Dinh THT, Kleimeier S. A credit scoring model for Vietnam's retail banking market. International Review of Financial Analysis, 2007, 16(5): 471–495. [doi: 10.1016/j.irfa.2007.06.001]

7 赖莹. 支持向量机在 P2P 借款人信用风险评估中的应用 [硕士学位论文]. 成都: 电子科技大学, 2018.

8 向晖, 杨胜刚. 基于多分类器组合的个人信用评估模型. 湖南大学学报 (社会科学版), 2011, 25(3): 30–33.
9 刘志惠, 黄志刚, 谢合亮. 大数据风控有效吗?——基于统计评分卡与机器学习模型的对比分析. 统计与信息论坛, 2019, 34(9): 18–26. [doi: 10.3969/j.issn.1007-3116.2019.09.003]
10 都红雯, 卢孝伟. 基于 SVM-Logistic 组合模型的 P2P 借款者信用风险评估——以微贷网为例. 生产力研究, 2018, (10): 31–36, 63. [doi: 10.3969/j.issn.1004-2768.2018.10.007]
11 谭中明, 谢坤, 彭耀鹏. 基于梯度提升决策树模型的 P2P 网贷借款人信用风险评测研究. 软科学, 2018, 32(12): 136–140.
12 费鸿雁, 黄浩. 基于模型融合的互联网信贷信用风险预测研究. 统计学与应用, 2019, 8(5): 823–834.
13 蒋翠清, 王睿雅, 丁勇. 融入软信息的 P2P 网络借贷违约预测方法. 中国管理科学, 2017, 25(11): 12–21.
14 祝钧桃, 姚光乐, 张葛祥, 等. 深度神经网络的小样本学习综述. 计算机工程与应用, 2021, 57(7): 22–33. [doi: 10.3778/j.issn.1002-8331.2012-0200]
15 Breiman L. Random forests. Machine Learning, 2001, 45(1): 5–32. [doi: 10.1023/A:1010933404324]
16 张国庆, 昌宁. 基于 LightGBM 的银行信用卡违约研究. 科技资讯, 2019, 17(12): 8–9.
17 沙靖岚. 基于 LightGBM 与 XGBoost 算法的 P2P 网络借贷违约预测模型的比较研究 [硕士学位论文]. 大连: 东北财经大学, 2017.
18 Ben-David S, Schuller R. Exploiting task relatedness for multiple task learning. Proceedings of the 16th Annual Conference on Learning Theory and 7th Kernel Workshop. Washington: Springer, 2003. 567–580.
19 Maghdid HS, Asaad AT, Ghafoor KZ, *et al*. Diagnosing COVID-19 pneumonia from X-ray and CT images using deep learning and transfer learning algorithms. Proceedings of SPIE 11734, Multimodal Image Exploitation and Learning 2021. Online: SPIE, 2021. 117340E.
20 Gao F, Yoon H, Wu T, *et al*. A feature transfer enabled multi-task deep learning model on medical imaging. Expert Systems with Applications, 2020, 143: 112957. [doi: 10.1016/j.eswa.2019.112957]
21 Dawud AM, Yurtkan K, Oztoprak H. Application of deep learning in neuroradiology: Brain haemorrhage classification using transfer learning. Computational Intelligence and Neuroscience, 2019, 2019: 4629859.
22 Bel N, Koster CHA, Villegas M. Cross-lingual text categorization. Proceedings of the 7th International Conference on Theory and Practice of Digital Libraries. Trondheim: Springer, 2003. 126–139.
23 Ling X, Xue GR, Dai WY, *et al*. Can Chinese web pages be classified with English data source? Proceedings of the 17th International Conference on World Wide Web. Beijing: ACM, 2008. 969–978.
24 龚澄. 迁移学习方法在个人信用违约预测中的作用 [硕士学位论文]. 成都: 西南财经大学, 2018.
25 戴文渊. 基于实例和特征的迁移学习算法研究 [硕士学位论文]. 上海: 上海交通大学, 2009.
26 Hinton GE, Osindero S, The YW. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527–1554. [doi: 10.1162/neco.2006.18.7.1527]
27 刘弘历, 武森, 魏桂英, 等. 基于深度神经网络的点击率预测模型. 工程科学学报, 2021, 2: 1–10.
28 徐桂琼, 李微. 基于组合分类的 P2P 贷款逾期风险预警研究. 管理现代化, 2019, 39(4): 9–12.

(校对责编: 孙君艳)