计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(3):338−344 [doi: 10.15888/j.cnki.csa.009003]

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

基于 CatBoost 用信预测模型的 TreeSHAP 解释性研究①

马 朔[1], 李 钊[2], 赵 军[3]

[1](宁夏大学 信息工程学院, 银川 750021)
[2](石嘴山银行股份有限公司 金融大数据实验室, 银川 750011)
[3](宁夏大学 前沿交叉学院, 中卫 755099)
通信作者: 赵 军, E-mail: tzhaoj@nxu.edu.cn

摘 要: 银行客户申请信用贷款在授信通过后, 精准预测客户是否用信及分析影响客户用信的关键因素, 对提高银行客户服务能力及盈利能力具有重要意义. 目前, 机器学习算法鲜有在用信预测方面的应用, 且金融用信领域缺乏模型可解释性的研究, 为此提出一种基于 CatBoost 的 TreeSHAP 解释性用信预测模型. 通过 CatBoost 构建用信预测模型, 利用 3 种超参数优化算法对该模型进行对比优化, 与基线模型在 4 项主要性能指标上进行实验对比, 结果表明经 TPE 算法优化后的模型性能均优于其他模型, 然后结合 TreeSHAP 方法从全局和局部的层面增强模型的可解释性, 解释性分析客户用信的影响因素, 为银行对客户进行精准化营销提供决策依据.

关键词: 用信预测; 可解释性; TPE; CatBoost; TreeSHAP; 机器学习

引用格式: 马朔,李钊,赵军.基于 CatBoost 用信预测模型的 TreeSHAP 解释性研究.计算机系统应用,2023,32(3):338–344. http://www.c-s-a.org.cn/1003-3254/9003.html

Research on Interpretative TreeSHAP Based on CatBoost's Credit Utilization Prediction Model

MA Shuo[1], LI Zhao[2], ZHAO Jun[3]

[1](School of Information Engineering, Ningxia University, Yinchuan 750021, China)
[2](Laboratory of Financial Big Data, Bank of Shizuishan, Yinchuan 750011, China)
[3](School of Advanced Interdisciplinary Studies, Ningxia University, Zhongwei 755099, China)

Abstract: It is essential for banks to accurately predict whether clients will use their credit and analyze key factors influencing credit utilization after these clients have been approved for credit, so as to improve their client service level and profitability. Currently, machine learning algorithms are rarely applied to credit utilization prediction, and there is a lack of research on model interpretability in the financial credit utilization field. Therefore, this study proposes an interpretative TreeSHAP credit utilization prediction model based on CatBoost. Specifically, a credit utilization prediction model is constructed by CatBoost and is compared and optimized by using three hyperparameter optimization algorithms. Then, the model is experimentally compared with baseline models in terms of four main performance metrics. The results show that the model optimized by the TPE algorithm outperforms other models. Finally, the interpretability of the model is enhanced locally and globally by the TreeSHAP method. Furthermore, factors influencing client credit utilization are interpretively analyzed, so as to provide a decision-making basis for banks to make accurate marketing to clients.

Key words: credit utilization prediction; interpretability; tree-structured parzen estimator (TPE); CatBoost; TreeSHAP; machine learning

① 基金项目: 国家自然科学基金 (71461025); 宁夏自然科学基金 (2020A1166)
收稿时间: 2022-08-17; 修改时间: 2022-09-15; 采用时间: 2022-09-27; csa 在线出版时间: 2022-11-29
CNKI 网络首发时间: 2022-11-30

线上个人信用贷款业务是商业银行直接面向个人客户的金融服务, 客户可以通过该项服务申请贷款, 但有些客户在完成贷款申请通过授信后, 并没有即时用信, 也就是没有发生用款行为, 有些客户在通过授信后马上就进行用信. 利用机器学习模型可以充分结合和利用银行金融大数据信息, 高效、批量地对客户是否用信进行预测, 但高性能机器学习模型由于其复杂的内部结构会引发不可解释的问题, 导致模型的输入和输出难以展现因果关系, 模型预测结果也难以被解释. 对于金融领域, 如果仅凭模型预测结果就轻易做出商业决策, 是具有很高的风险性的. 而模型可解释性方法可以结合实际业务理解对机器学习模型预测结果进行解释性分析, 使预测结果符合业务认知, 更加有理有据. 机器学习模型可解释性方法可以满足商业银行对金融业务场景下的模型可解释性要求, 推动高性能机器学习模型在金融领域的应用. 本文研究的主题就是根据客户相关信用信息, 预测客户在通过授信后是否用信, 并重点研究结合可解释性方法分析影响客户用信的关键因素.

目前, 国内外相关研究领域对客户用信预测的研究较少, 金融用信方面模型可解释性的研究不足. 在客户用信预测的相关研究工作中: 倪政[1] 提出基于随机森林的农户用信行为预测模型, 并使用 LIME 可解释性框架分析影响农户是否用信的关键因素, 为银行对信贷农户采用不同营销策略提供依据, 预测结果的相关性能指标优于神经网络模型及支持向量机模型, 但在结合算法模型时没有做过多的优化策略. 在模型可解释性的相关研究工作中: 雷欣南等[2] 结合 SHAP 解释方法提出小微企业违约识别的机器学习模型, 将 SHAP 方法应用于实际业务中, 分析比较了导致小微企业发生违约的关键特征. 蔡青松等[3] 提出基于可解释集成学习的信贷违约预测, 通过模型融合方法, 引入 LIME 解释融合模型的预测结果, 提高了信贷违约预测的精确性和可解释性. 孔令莹[4] 提出基于 LightGBM 结合 SHAP 方法建立客户消费信贷违约风险评估模型, 分析了对信贷违约有关键影响的不同因素以及特征变化对违约风险预测的影响. Chen 等[5] 提出了一种两层加性信用风险可解释模型, 该方法提供了 3 种类型的解释, 具有全局一致性, 有助于特征重要性分析及解释.

上述已有的相关研究工作中, 很少有在客户用信预测方面的研究, 且在模型可解释性方面的研究中, 虽然采用了各种集成学习算法提高了模型预测的准确性, 但所建立的模型难以平衡精度和可解释性, 并且缺乏足够的可解释性, 而商业银行以及其他金融机构对金融业务场景下的模型可解释性的要求很严格. 为了解决以上研究中存在的问题, 在平衡模型预测精度和可解释性的基础上, 本文提出使用 CatBoost 模型进行客户用信行为的预测, 并引入 TreeSHAP 方法研究模型的可解释性. 本文主要贡献在于: 通过 TreeSHAP 方法在一定程度上克服了 CatBoost 模型的黑箱特性, 对影响客户是否用信的关键因素结合金融业务场景进行可解释性分析, 不仅可以实现在总体层面上解释某些因素如何影响客户用信, 也可以实现在单个样本层面上解释哪些因素的变化会导致客户用信. 进行客户用信预测可解释性研究, 为商业银行精准化营销提供依据, 具有很大的应用价值.

1 相关理论概述

1.1 CatBoost 模型

CatBoost[6] 是一种以对称决策树为基学习器的 GDBT 算法框架下的一种改进实现算法, 主要优势是能够对类别特征进行高效处理、解决了预测偏移以及梯度偏差的问题, 从而减少过拟合, 进而提高模型的泛化能力和准确性. CatBoost 处理离散特征的主要方法是 Greedy TS, 其思想是在决策树中, 将标签平均值作为节点分裂的标准, 公式如下:

$$x_{i,j}=\frac{\sum_{k=1}^{n}[x_{k,j}=x_{i,j}]\cdot Y_k+\alpha\cdot P}{\sum_{k=1}^{n}[x_{k,j}=x_{i,j}]+\alpha} \tag{1}$$

其中, $x_{i,j}$ 表示第 j 个特征的第 i 类值, Y_k 表示所对应的标签值, 分子表示第 j 个特征的第 i 类值所对应的标签值的总和, 分母表示第 j 个特征的第 i 类值的数量. P 表示添加的先验项, α 是大于 0 的权重系数. 添加先验分布项, 可以减少数据噪音, 克服低频类别数据对于模型的影响.

1.2 TreeSHAP 解释性方法

集成学习模型自有的特征重要性评估方法只能表示出哪个特征重要, 并不能解释该特征如何影响模型的预测结果, 而 SHAP 事后可解释方法不但能反映出每一个样本中的特征对预测结果的影响作用, 并且还

能表现出这种作用的正负性. SHAP (Shapley additive explanation) 是由 Lundberg 等[7] 提出的一种对模型进行解释的方法. SHAP 基于 Shapley 值, 将该值解释为一种加性特征归因方法 (addititve feature attribution method), 将模型的预测值解释为二元变量的线性函数, 函数公式如下:

$$g(z) = \phi_0 + \sum_{k=1}^{M} \phi_k \tag{2}$$

其中, g 表示解释模型, M 表示输入特征的数量, ϕ_0表示一个常数, ϕ_k表示每个特征 k 的 Shapley值, 公式如下:

$$\phi_k = \sum_{N \subseteq \{M \setminus x_k\}} \frac{(|M| - |N| - 1)!|N|!}{|M|!} \{f(x_{N \cup \{k\}}) - f(x_N)\} \tag{3}$$

其中, M 表示特征集合, N 表示$\{M\setminus x_k\}$的子集合, 分式表示不同特征组合对应的概率, $f(x_{N\cup\{k\}})$ 与 $f(x_N)$ 分别表示不同特征组合下 x_k 入模与不入模时模型的预测结果.

针对树集成模型, Lundberg 等[8] 提出了 SHAP 的改进方法 TreeSHAP, 该方法运算速度快、不需要抽样、通过树模型中的节点来计算 Shapley 值, 尤其适用于基于决策树的集成学习模型. TreeSHAP 的基本计算原理如下: 给定一棵树, 针对某个样本 s 和特征子集 T, 若子集 T 含有全部特征, 则叶子结点的预测值就是该特征下的模型的输出值; 若子集 T 为空, 则将全部叶子结点的预测值的加权平均作为输出值; 若子集 T 包含部分特征, 则删除由于去掉部分特征后无法通达的叶子结点, 在剩下的结点中取加权平均作为输出值. 通过 TreeSHAP 方法, 可以优化计算 Shapley 值的算法从而显著缩短计算时间、可以将局部解释直接扩展到抓取模型内部的交互效应、可以计算出全部样本集中对应某个特征的 Shapley 值, 将它们的均值作为该特征的重要性值, 从而基于众多的局部解释得到模型的全局解释. 接下来的实验部分, 将使用 TreeSHAP 方法来解释 CatBoost 模型.

2 特征工程

2.1 数据预处理

本文所采用的数据集, 来自 S 银行 2021 年 4 月至 2022 年 4 月期间线上个贷业务经 MD5 加密后的客户用信、授信和基本信息数据, 对某些数据特征名进行模糊处理, 然后进行数据集成并去重, 删除一些无关特征后, 得到整合信用数据集, 共有 158 个特征、91 448 条样本. 进行缺失值处理, 删除数据集行缺失率大于 0.5 的 2 579 个样本, 没有列缺失率大于 0.9 的特征, 无需删除特征. 对有缺失值的样本根据 S 银行提供的特征缺失值填充规则进行填充. 进行异常值处理, 利用四分位距原理 (IQR)[9] 对异常值进行检测, IQR 表示上四分位与下四分位的差值, 超过 IQR 规定上下界以外的点为异常值. 在本文数据集中, 存在异常值的样本量相对总体样本量较少, 直接删除存在异常值的样本, 不会对整体样本量产生较大的影响.

2.2 特征选择

选择合适的特征及特征数可有效防止过拟合问题, 第 1 步进行基于 IV 值[10] 的初步选择, IV 值表征输入特征对目标特征的预测能力, 公式如下:

$$IV = \sum_{i}^{m} (P_{yi} - P_{ni}) \ln\left(\frac{P_{yi}}{P_{ni}}\right) \tag{4}$$

其中, i 表示组数, m 表示在该特征上样本划分的总组数, P_{yi}、P_{ni} 分别表示在第 i 组中用信占总用信和不用信占总不用信的样本比例. 删除 IV 值小于 0.02 的无预测能力的特征, 选择了 86 个特征. 第 2 步进行基于 Pearson 相关系数[11] 的特征选择, 去除特征冗余, 公式如下:

$$\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{\sum[(x-\bar{x})(y-\bar{y})]}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \tag{5}$$

其中, $Cov(x, y)$ 表示特征 x、y 的协方差, σ_x 和 σ_y 表示两个特征的标准差. 当两个特征相关系数的绝对值大于 0.7 时, 表示这两个特征强相关, 则删除其中 IV 值较小的特征, 进一步选择了 34 个特征.

3 实验设计与分析

整个实验流程如图 1 所示, 主要包括特征工程、基于 CatBoost 构建用信预测模型、3 种超参数优化算法对比分析、4 项模型性能指标对比分析、TreeSHAP 模型可解释性分析.

3.1 模型评估指标

实验采用精确率 (*Precision*)、召回率 (*Recall*)、准确率 (*Accuracy*)、AUC 值这 4 项分类评估指标对模型性能进行对比评估. AUC 值是接收者工作特征 (ROC) 曲线的量化表示, AUC 值越接近于 1, 模型预测性能

越优. 前 3 个指标的公式如下所示, 混淆矩阵如表 1 所示.

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

$$Accuracy = \frac{TP+TN}{FN+FP+TP+TN} \tag{8}$$

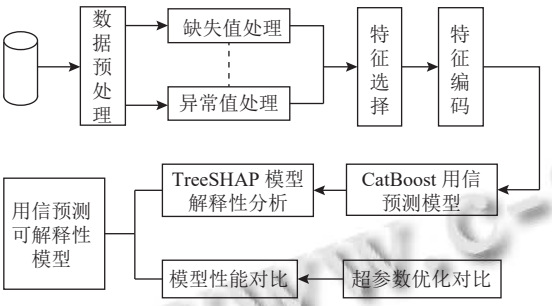

图 1 实验流程

表 1 混淆矩阵

真实结果	预测结果	
	用信	未用信
用信	*TP*	*FN*
未用信	*FP*	*TN*

3.2 模型超参数优化对比分析

实验分别利用网格搜索 (GridSearch)、随机搜索 (RandomSearch) 和 TPE 算法[12] 对 CatBoost 模型的 4 种核心参数进行超参数优化, 以 AUC 值作为调参优化指标. 超参数调优结果对比如表 2 所示, 调优算法性能指标对比结果如表 3 所示. 由表 3 可知, 经过 TPE 算法优化后的模型 AUC 值最高.

表 2 超参数调优结果对比

参数	默认值	RandomSearch	GridSearch	TPE
learning_rate	0.03	3	0.1	0.06
depth	6	10	6	6
l2_leaf_reg	3	4	9	6
min_data_in_leaf	20	12	10	13

表 3 调优算法性能指标对比

调优算法	*Precision*	*Recall*	*Accuracy*	AUC
CatBoost	0.6844	0.5564	0.8087	0.8535
R-CatBoost	0.6762	0.5436	0.7877	0.8536
G-CatBoost	0.6849	**0.5570**	0.8028	0.8538
T-CatBoost	**0.6875**	0.5527	**0.8092**	**0.8541**

3.3 模型性能对比分析

实验将优化后的 T-CatBoost 模型与未优化的 CatBoost 模型以及其他 5 种主流机器学习模型的预测结果性能指标进行对比分析, 结果如表 4 所示, ROC 曲线对比如图 2 所示.

表 4 模型性能指标对比

模型	*Precision*	*Recall*	*Accuracy*	AUC
LR	0.6267	0.4499	0.7720	0.8068
AdaBoost	0.6440	0.4963	0.7874	0.8230
RandForest	0.6881	0.4724	0.7973	0.8337
XGBoost	0.6816	0.5202	0.8025	0.8449
LightGBM	0.6771	0.5511	0.8056	0.8505
CatBoost	0.6844	**0.5564**	0.8087	0.8535
T-CatBoost	**0.6875**	0.5527	**0.8092**	**0.8541**

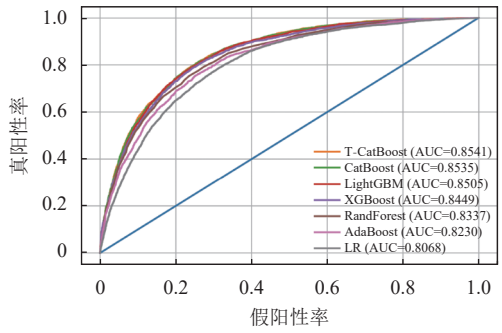

图 2 ROC 曲线对比

由表 4 及图 2 的实验结果可知, 经 TPE 算法超参数调优后的 CatBoost 模型在精确率、准确率和 AUC 值上的性能表现均优于其他模型. CatBoost 模型的优势在于对离散型特征采用目标统计的编码方式, 可以有效降低数据噪音, 并且由于解决了梯度偏差和预测偏移的缺陷, 从而有效减少了模型训练过拟合问题, 提高了模型预测的性能.

4 TreeSHAP 模型解释性分析

4.1 全局可解释性分析

从全局层面对用信预测模型进行解释与分析, 图 3 对影响客户是否用信的特征按重要性进行排序, 显示了对客户是否用信影响作用最大的 12 个特征. 特征 feature1 表示“贷记卡额度使用率”, 该特征是影响客户是否用信的最重要因素, 贷记卡额度使用率越高, 则客户用信的概率越大. 而对于授信额度特征, 其 SHAP 值大量聚集于平均值以下, 属于左偏分布, 说明其授信额度越低, 客户用信概率越小.

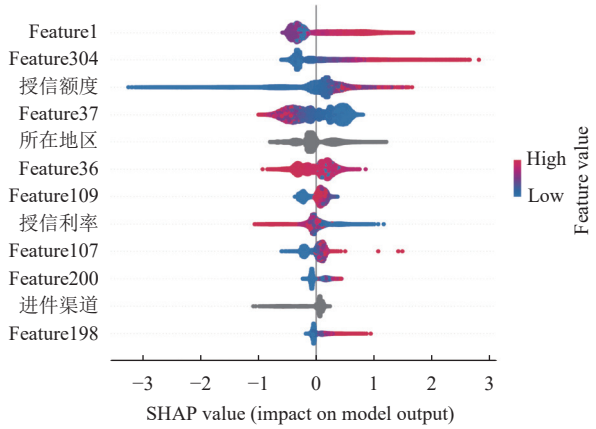

图 3 SHAP 特征概要

4.2 局部可解释性分析

从局部层面对单个客户是否用信的影响因素进行解释性分析, 图 4(a) 为某一预测结果为用信客户的特征贡献, 图 4(b) 为某一预测结果为不用信客户的特征贡献, 其中红色部分表示对预测为用信有正向影响的特征, 会增大模型将该客户判断为用信客户的概率, 蓝色部分表示对预测为用信有负向影响的特征, 会增大模型将该客户判断为不用信客户的概率. 白色箭头表示特征之间的分割线, 相邻分割线之间的距离, 则表示所对应特征及其特征值对预测结果的影响程度, 距离越长, 则影响程度越大.

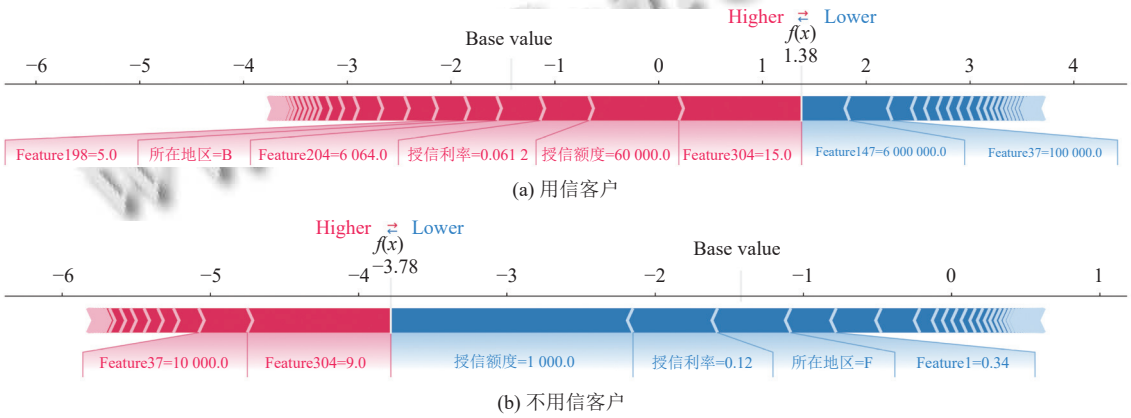

图 4 SHAP 特征贡献

图 4(a) 中, 该客户样本的预测概率值 $f(x)$ 为 1.38, 远大于整个数据集的基准值 (base value), 表示通过 TreeSHAP 方法解释的 CatBoost 用信预测模型预测该客户很有可能会用信, 被预测为用信客户的最大影响因素是 feature304, 其值为 15, 其他主要影响因素为授信额度较高为 60 000 元、授信利率较低为年化 6.12% 等因素, 特征 feature304 表示“近 12 个月贷款审批查询次数”.

图 4(b) 中, 该客户样本的预测概率值 $f(x)$ 为−3.78, 远小于整个数据集的基准值 (base value), 表示通过 TreeSHAP 方法解释的 CatBoost 用信预测模型预测该客户很有可能会不用信, 被预测为不用信客户的最大影响因素是授信额度较低只有 1 000 元, 其他主要影响因素是授信利率较高为年化 12.0%、feature1 为 0.34 等因素, 这就导致了客户最终可能会选择不用信.

4.3 特征依赖性分析

分析单个特征如何影响用信预测模型的预测结果, 选择 feature26 和授信额度这两个特征分别绘制 SHAP 特征部分依赖图. 如图 5(a) 所示, 依赖图的横轴表示所选特征的值, 纵轴表示所选特征的 SHAP 值, 右边第 3 轴表示与所选特征交互作用最强的特征的值, 该特征是算法自动选择的, 散点颜色越红, 表示该特征的值越大. 随着 feature26 的增大, SHAP 值呈现先降低再升高的变化趋势, 表明 feature26 增大到某一阈值之前对客户用信存在反向作用, 增大到某一阈值之后对客户用信存在正向作用, feature26 表示“信用记录月份数”. 图 5(b) 中, 随着授信额度的增大, SHAP 值在增加并趋于稳定, 表明授信额度提高对客户用信存在正向作用, 但授信额度提高到 25 000 元之后, 红色散点整体高于蓝色散点, 且所有样本的 SHAP 值显现出水平分布的情况, 此时授信额度已不再是影响客户用信的主要因素, 其对客户用信的影响就不再发生显著变化. 在相同的授信额度条件下, 特征 feature37 的值越大, 预测为用信客户的概率越高.

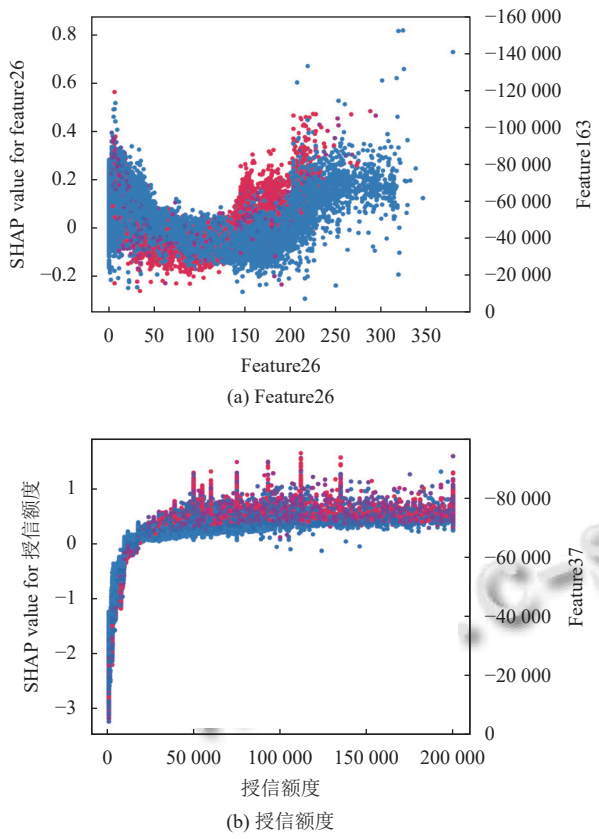

图 5 SHAP 特征依赖

4.4 特征重要性分析

对比分析不同模型和 TreeSHAP 方法的特征重要性排序, 可综合评估出影响客户是否用信的关键因素. 由图 6 可知, TreeSHAP 方法、CatBoost 模型、LightGBM 模型、XGBoost 模型给出的特征重要性排序并不完全相同, 综合分析 4 种排序结果, 可知对客户是否选择用信具有显著性影响的因素: feature1、feature304、授信额度、feature37、所在地区、授信利率和 feature36 等. 在实际业务中, 需要重点关注这一组特征及其数据分布情况.

5 结论与展望

客户用信率的提高对商业银行进行精细化运营和精准化营销具有重要作用, 准确预测客户是否用信且分析客户用信的关键影响因素则能有效提高客户服务质量和效率. 本文提出一种基于 CatBoost 的客户用信预测模型, 并结合 TreeSHAP 方法对用信预测模型进行解释性分析. 同时, 通过 3 种超参数优化方法对比调参进一步提升了模型的性能, 与逻辑回归、AdaBoost、随机森林、LightGBM 和 XGBoost 这 5 种主流机器学习模型进行对比实验, 证明了本文所提模型的优越性. 最后, 通过 TreeSHAP 方法结合实际金融业务对模型进行全局和局部解释性分析以及特征依赖性分析, 并通过对比不同模型的重要性排序, 综合分析出了影响客户是否用信的关键因素.

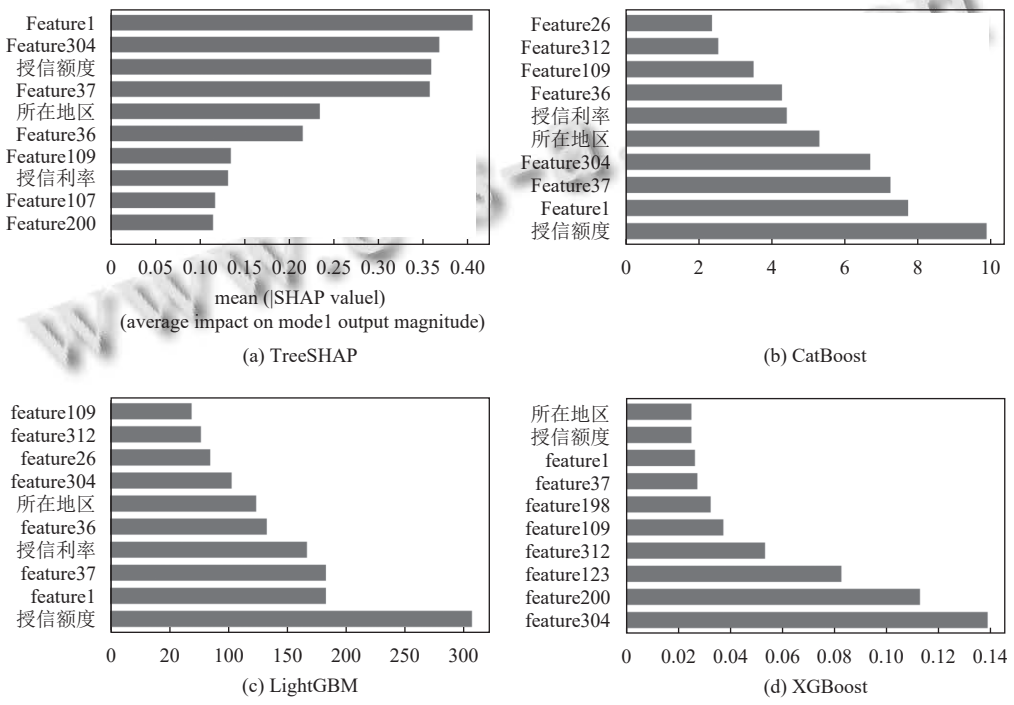

图 6 特征重要性 (其中, 横坐标表示模型自有特征重要性评估方法的分数值, 纵坐标表示特征名)

由于本文数据集来自商业银行真实信贷业务数据, 目前缺少公开的客户用信数据集作为对照, 下一步工作主要是在公开数据集上进行模型的迁移实验, 以及考虑更加有效的模型优化方法和模型解释方法, 进一步提高模型的精度、可解释性及其实际应用价值.

参考文献

1 倪政. 基于随机森林的兴农卡农户用信预测模型及应用研究 [硕士学位论文]. 武汉: 中南林业科技大学, 2019.

2 雷欣南, 林乐凡, 肖斌卿, 等. 小微企业违约特征再探索: 基于 SHAP 解释方法的机器学习模型. 中国管理科学, 2022: 1–13. [doi: 10.16381/j.cnki.issn1003-207x.2021.0027]

3 蔡青松, 吴金迪, 白宸宇. 基于可解释集成学习的信贷违约预测. 计算机系统应用, 2021, 30(12): 194–201. [doi: 10.15888/j.cnki.csa.008220]

4 孔令莹. 基于 TPE-LightGBM 算法和 SHAP 值的信贷违约预测 [硕士学位论文]. 湘潭: 湘潭大学, 2021.

5 Chen CF, Lin KC, Rudin C, *et al*. An interpretable model with globally consistent explanations for credit risk. arXiv:1811.12615, 2018.

6 Prokhorenkova LO, Gusev G, Vorobev A, *et al*. CatBoost: Unbiased boosting with categorical features. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 6639–6649.

7 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 4768–4777.

8 Lundberg SM, Lee SI. Consistent feature attribution for tree ensembles. arXiv:1706.06060, 2017.

9 Vinutha HP, Poornima B, Sagar BM. Detection of outliers using interquartile range technique from intrusion dataset. Information and Decision Sciences: Proceedings of the 6th International Conference on FICTA. Singapore: Springer, 2018. 511–518.

10 Freedman S, Jin GZ. The information value of online social networks: Lessons from peer-to-peer lending. International Journal of Industrial Organization, 2017, 51: 185–222. [doi: 10.1016/j.ijindorg.2016.09.002]

11 Ly A, Marsman M, Wagenmakers EJ. Analytic posteriors for Pearson's correlation coefficient. Statistica Neerlandica, 2018, 72(1): 4–13. [doi: 10.1111/stan.12111]

12 Erwianda MSF, Kusumawardani SS, Santosa PI, *et al*. Improving confusion-state classifier model using XGBoost and tree-structured parzen estimator. 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). Yogyakarta: IEEE, 2019. 309–313.

(校对责编: 牛欣悦)