李春雷,王文生,郭雷风,等. 基于集成学习算法的返贫人口识别模型——以 H 省 F 县贫困户建档立卡数据为例[J]. 江苏农业科学,2021,49(17):231 – 237.
doi:10.15889/j.issn.1002 – 1302.2021.17.041

基于集成学习算法的返贫人口识别模型
——以 H 省 F 县贫困户建档立卡数据为例

李春雷[1], 王文生[1,2], 郭雷风[1], 陈桂鹏[3]
(1. 中国农业科学院农业信息研究所,北京 100081; 2. 国家农业农村部信息中心,北京 100020;
3. 江西省农业科学院农业经济与信息研究所,江西南昌 330200)

摘要:2020 年底精准扶贫工作胜利完成,但绝对贫困和区域性整体贫困的消除并不意味着贫困的消失和扶贫工作的结束。党中央多次强调要健全防止返贫动态监测和帮扶机制,对易返贫致贫人口实施常态化监测。当前对返贫动态监测的研究多为宏观政策性内容,对贫困人口进行返贫识别的微观操作性研究较少。针对上述问题,利用贫困户建档立卡数据进行数据处理选取 14 维特征,构建基于集成学习算法的返贫人口识别模型进行贫困人口分类。结果表明,经调优的 XGBoost 算法模型取得最优结果,对已脱贫、未脱贫及返贫 3 类人员分别达 97.43%、92.44%、97.04% 的识别准确率,总体达到 96.81% 的准确率,能够较好识别出贫困人口贫困类别。为帮扶工作人员的防返贫动态监测和帮扶工作提供技术支持。

关键词:建档立卡;集成学习;返贫识别;动态监测
中图分类号: F323.8;TP181　　**文献标志码**: A　　**文章编号**:1002 – 1302(2021)17 – 0231 – 07

在 2020 年 12 月 3 日中共中央政治局常务委员会会议上,中共中央总书记习近平宣布,经过 8 年持续奋斗,现行标准下农村贫困人口全部脱贫,消除了绝对贫困和区域性整体贫困,取得了脱贫攻坚重大胜利。随着精准扶贫的完成,全国约 9 900 万贫困人口实现脱贫,贫困地区的已脱贫贫困人员返贫问题也随之显现。2020 年以来受极端气候灾害、新冠疫情等突发事件以及国际形势变化的影响,已脱贫人口面临较大的返贫压力,以及部分边缘人口也面临致贫风险。因此,“后扶贫时代”的关注焦点是怎样实现可持续脱贫。党的十九大明确,农村绝对贫困人口实现脱贫,并不意味着农村贫困的消失和扶贫工作的结束,要进一步巩固建设成果,防止返贫。

现阶段对防止返贫监测预警的研究多为政策干预层面,如根据多维指标建立评价体系进行相对贫困预警监测分级,采取分级治理措施[1]。而对于返贫人口的识别监测工作的具体操作研究较少,主要工作方式仍是依赖精准扶贫阶段建设的扶贫工作体系,扶贫干部、信息员等一线扶贫工作者入户

收稿日期:2021 – 02 – 02
基金项目:江西现代农业科研协同创新专项(编号:JXXTCX201801 – 03)。
作者简介:李春雷(1994—),男,河北邢台人,硕士研究生,主要从事信息技术农业应用相关研究。E – mail:lcl050024@126.com。
通信作者:王文生,博士,研究员,博士生导师,主要从事农业信息化相关研究。E – mail:13911359883@163.com。

Plant Physiology,1987,84(2):450 – 455.
[12]蒋跃明. 荔枝采后果皮中巯基化合物含量变化及与褐变的关系(简报)[J]. 植物生理学通讯,1997,33(4):262 – 264.
[13]孙亚莉,徐庆国,贾　巍. 镉胁迫对水稻的影响及其调控技术研究进展[J]. 中国农学通报,2017,33(10):1 – 6.
[14]章秀福,王丹英,储开富,等. 镉胁迫下水稻 SOD 活性和 MDA 含量的变化及其基因型差异[J]. 中国水稻科学,2006,20(2):194 – 198.
[15]韩淑梅,陈贵川,侯双双,等. 孔雀草体内低分子质量巯基化合物对重金属镉的响应[J]. 种子,2018,37(10):36 – 40.
[16]王凯荣,张玉烛,胡荣桂. 不同土壤改良剂对降低重金属污染土壤上水稻糙米铅镉含量的作用[J]. 农业环境科学学报,2007,26(2):476 – 481.
[17]丁凌云,蓝崇钰,林建平,等. 不同改良剂对重金属污染农田水稻产量和重金属吸收的影响[J]. 生态环境,2006,15(6):1204 – 1208.
[18]Yin X L,Xu Y M,Huang R,et al. Remediation mechanisms for Cd – contaminated soil using natural sepiolite at the field scale[J]. Environmental Science: Processes & Impacts,2017,19(12):1563 – 1570.

摸排进行信息采集和回访,将入户结果整理后自下向上层层上报[2]。在 2020 年 12 月 28 日中央农村工作会议上,党中央决定从脱贫之日起设立 5 年过渡期,过渡期内要保持主要帮扶政策总体稳定,逐步实现向全面推进乡村振兴平稳过渡。这个过程中,扶贫工作队以及各单位抽调的帮扶人员必然要逐步撤出,原有的扶贫工作机制必然要有所转变。加强对大数据等信息技术的利用,是实现对重点人群常态化监测的必然要求,也是减轻扶贫工作人员工作压力、提高返贫监测和帮扶工作效率的重要保障。

近几年大数据、机器学习等技术也开始被应用于扶贫工作中。在贫困人口精准识别工作中利用随机森林算法对贫困人口进行精准识别能够取得不错的效果[3],但相关工作多采用社会科学调查数据,存在成本较高、周期较长的不足。部分研究人员提出利用大数据信息系统进行返贫预警[4],但是对如何利用大数据进行返贫预警的操作多为宏观阐述。国外学者在研究减贫问题过程中提出利用深度学习技术基于低成本高分辨微星图像估计区域财富和消费水平,以此弥补缺乏大规模可靠公共数据的缺陷[5]。

自精准扶贫工作开展以来,在中央和地方共同努力下,各地针对本地区贫困户进行了建档立卡等多方面数据采集工作[6],积累了大量的能够反映区域特征的贫困人口数据。基于现有的大规模、细粒度的数据优势,深入挖掘利用建档立卡数据,以此提升精准识别精度、为帮扶政策制定提供决策依据。有研究者利用机器学习算法结合建档立卡数据进行帮扶方式推荐[7],为扶贫工作者提供扶贫方式参考。而当前对挖掘到的建档立卡数据进行返贫识别的研究较少。本研究利用精准扶贫工作中积累的建档立卡数据,采用能够处理多数据类型、训练速度快、鲁棒性较强的 XGBoost 等集成学习算法建模,对贫困人口进行已脱贫、贫困、返贫三分类识别,对已脱贫人口长期跟踪,对返贫贫困人口动态监测和及时干预,减轻扶贫工作人员工作压力,提高工作效率,使精准扶贫已取得的工作成果得到保障。

1　数据来源及处理

1.1　数据来源及试验设备

本研究所用的数据来源于笔者所在团队对国家级贫困县 H 省 B 市 F 县建设的帮扶项目“精准扶贫大平台”,该项目旨在从全要素、全生命周期提升当地的精准扶贫工作的信息化水平,项目建设期间帮助 F 县当地各单位帮扶人员帮扶信息系统,利用 web 端平台、手机 app 等多种方式助力 F 县精准扶贫工作,提升精准扶贫工作效率。所用的数据节点为 2020 年初 F 县贫困人员信息,共计贫困户 31 438 户 92 482 人,其中尚未脱贫 11 367 人,已脱贫 79 777 人,返贫人员 1 338 人。

所用开发语言为 Python 3.7 配合 sklearn 工具包和 XGBoost、LightGBM 及 CatBoost 对应的 Python 工具包;所有计算运行环境均为 Win10 操作系统,采用 i5 - 9600KF 6Core 处理器。

1.2　原始数据处理及特征构建

贫困人员家庭人均纯收入是其一段时期内的收入反映,但这一指标容易受到短期帮扶政策因素或者贫困人员家庭变故的影响。故仅凭借收入这一单独指标来认定贫困人口脱贫状态存在一定的局限性,在当下以及日后的扶贫以及防止返贫工作中是远远不够的。运用多维贫困测度方法,从多个维度对贫困人口进行识别,更加精准地发现贫困人口困难所在,有针对性进行帮扶,对贫困人口脱贫动态追踪管理,才能够有效提升精准扶贫效率[8]。

根据罗丽在可持续升级分析框架的基础上构建的多维贫困识别指标体系中的指标[3],从劳动能力、教育文化、劳动技能、基础设施、家庭收入等方面对建档立卡数据进行格式统一整理、转化和清洗,对收集到的贫困人口原始数据进行处理和特征筛选。各地大量记录表明,疾病医疗是导致贫困或返贫的重要原因之一[9]。许多原本依靠自身务工摆脱贫困的家庭,由于家庭成员患病失去务工收入且还有可能需要家中其他劳动力辞工照顾,使得本已脱贫的家庭再度返贫。故据此增加构建家庭疾病人口比率这一特征,及根据户号、家庭人口数量以及人员健康情况信息计算患病和残疾人数占家庭人口总数比例。为便于建模分析,将原始数据中的各项贫困特征类别数据转化为数值数据,具体数值定义见表 1。

原始数据中的每户人数、外出务工时间(月)、平均收入 3 项数值型数据均不做转换处理。根据表中数值定义将贫困人口原始数据转换为数值型,转换过后贫困人口数据转换成为一个贫困数据矩阵,即可作为算法的输入数据进而构建贫困人口分类

表 1　贫困人口类别特征数据数值定义

特征	含义	数值定义
贫困类型	对贫困人口类别划分	0 = 已脱贫;1 = 未脱贫;2 = 返贫
住址	贫困人口所住村庄	根据数据中出现先后顺序共计 0 ~ 207 类,208 个村庄
民族	贫困人口所属民族	根据数据中出现先后顺序共计 0 ~ 10 类,11 个民族
文化程度	对贫困人口学历水平划分	文盲半文盲 = 0;在读 = 1;小学 = 2; 初中 = 3;高中 = 4; 大专 = 5;本科及以上 = 6
健康状况	对贫困人口身体健康水平划分	健康 = 0;多种疾病 = 1;患有大病 = 2;残疾 = 3;长期慢性病 = 4
劳动能力	对贫困人口劳动能力水平划分	0 = 弱劳动力或半劳动力;1 = 技能劳动力;2 = 无劳动力;3 = 普通劳动力
大病医疗	贫困人口是否参加大病医疗	否 = 0;是 = 1;
危房户	贫困人口是否居住危房	否 = 0;是 = 1;
致贫原因	贫困人口致贫原因划分	交通条件落后 = 0;其他 = 1;因丧 = 2;因婚 = 3;因学 = 4;因残 = 5;因灾 = 6;因病 = 7;缺劳力 = 8;缺土地 = 9;缺技术 = 10;缺水 = 11;缺资金 = 12;自身发展动力不足 = 13

算法模型,最终构建成为包含 14 个特征 92 482 条数据的数据集。利用 sklearn 工具包的数据划分工具,将数据随机打乱,根据类别比例,80% 划分为训练集,其他用作验证集。

2　模型介绍

本研究采用近年来在实际业务场景中有优异表现的集成学习算法来构建贫困人口识别模型。集成学习即使用一系列的学习器进行学习,采用某种规则将得到的学习器的学习结果进行整合,从而得到更好的学习效果。

2.1　极端梯度提升(XGBoost)

XGBoost 是由陈天奇博士团队 2014 年开源的机器学习项目,在 2016 年机器学习比赛中大放异彩,之后便成为了各类比赛的首选算法[10]。XGBoost 的目标函数:

$$L(\varphi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)。 \quad (1)$$

相比于原始 GBDT,多了正则项,能够减少过拟合的可能,同时加快了收敛速度。

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \| w \|^2。 \quad (2)$$

式中:γ 表示树分裂难度系数,来控制树的生成;T 表示叶子节点个数;λ 表示的是 L2 正则系数,如此对叶子节点个数进行惩罚,相当于在训练过程中剪枝。将损失函数用泰勒公式二阶展开,如此新的目标函数能够取得更快的收敛速度和准确性,最终目标函数变为公式(3)。

$$obj^{(i)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T。 \quad (3)$$

式中:$\boldsymbol{I}_j \subset \{q(X_i) = j\}$ 表示该树中索引为 j 的叶子上含有的样本集合,在 XGBoost 中用 $q(x_i)$ 表示样本 x_i 输入到模型后会被划分到哪个叶子节点 h_i 为损失函数 $L(\varphi)$ 的二阶导数;g_i 为损失函数 $L(\varphi)$ 的一阶导数。

2.2　LightGBM

LightGBM 为 2017 年微软亚洲研究院开源的模型[11],是在 XGBoost 上进一步改进的,也是基于 GBDT 算法演变而来的。XGBoost 在选择最优分裂点时需要扫描每一个样本点的特征,非常耗费时间和内存。LightGBM 主要解决了 GBDT 在大数据情况下的问题,让 GBDT 更方便用于实践。LightGBM 采用 histogram 算法,将样本浮点特征离散化,进行分桶形成 K 个整数特征,同时构造宽度为 K 的直方图。在遍历同时,将离散值作为累计索引进行统计,根据离散值寻找最佳分割点。利用直方图做差加速,将原本需要遍历叶子节点所有数据简化为遍历直方图的 K 个桶。LightGBM 使用带有深度限制的按叶子生长(leaf - wise)算法,更加高效。每次从当前所有叶子中,找到分裂增益最大的叶子进行分裂,如此循环。在分裂次数相同的情况下,leaf - wise 可以降低更多误差,取得更好的精度。防止产生较深的决策树,出现过拟合,LightGBM 增加了一个最大深度限制用来防止过拟合。

2.3　CatBoost

CatBoost 同样在 2017 年由俄罗斯的搜索引擎公司 Yandex 的研究团队提出的一种基于 boosting 的算法[12]。其对类别特征有着很好的支持。一般的梯度提升算法,最常用的是将类别特征转换为数值型来处理,类别数量差异较大时,这种做法容易产生过拟合。CatBoost 给出一种解决方案,可以减少过拟合发生。首先对所有样本进行随机排序,原顺序为 $\boldsymbol{c} = (c_1, \cdots, c_n)$,产生 $\boldsymbol{c}$ 的一次随机遍历序

列，用遍历的前 p 个记录针对类别型特征中的某个取值，每个样本的该特征转为数值型时都是基于排列在该样本之前的类别标签取均值，同时加入先验值 P 和参数 $\alpha>0$ 来控制低频类别噪音，公式如下：

$$\frac{\sum_{j=1}^{p}\left[x_{j,k}=x_{i,k}\right]\cdot Y_i+\alpha\cdot P}{\sum_{j=1}^{n}\left[x_{j,k}=x_{i,k}\right]+\alpha}。\tag{4}$$

CatBoost 采用排序提升（ordered boosting）的方式替换传统 GDBT 算法中的梯度计算方法，能够减小梯度估计偏差，提升模型泛化能力。

3　结果与分析

3.1　评价指标

混淆矩阵（confusion matrix）是评价模型精度的标准格式，使用 n 行 n 列的矩阵形式表示。矩阵每一列代表预测值，每一行代表实际值（表 2）。它的作用是表明每个类别之间是否有混淆，也就是模型到底判断对了多少个结果，判断错了多少个结果。同时混淆矩阵也能够帮助理解准确率、精确率和召回率，并利用 F_1 值综合衡量精确率和召回率。

表 2　多分类混淆矩阵

类别	预测结果		
	类别 0	类别 1	类别 2
0（已脱贫）	T_0	F_{10}	F_{20}
1（未脱贫）	$F_{0.1}$	T_1	F_{21}
2（返贫）	$F_{0.2}$	F_{12}	T_2

注：$T_i(i=0,1,2)$ 表示第 i 分类正确的样本数量，$F_{ij}(i,j=0、1、2)$ 表示实际为 i 类被错分为 j 类的样本数量。

正确率 $=(\sum T_i)/n(i=0、1、2,n$ 为样本总量）；

误差率 $=(\sum F_{ij})/n(i=0、1、2,n$ 为样本总量）；

精确率 $(P_i)=(T_i)/(T_i+F_{ji})(i,j=0、1、2,j\neq i)$；

召回率 $(R_i)=(T_i)/(T_i+F_{ij})(i,j=0、1、2,j\neq i)$；

$F_1(f_i)=2P_iR_i/(P_i+R_i)$。

3.2　模型结果比较

利用 3 种算法 XGBoost、LightGBM、CatBoost 构建迭代次数 1 500 次，其余参数默认的基线模型，比较基线模型初步结果（表 3）。

将 3 个模型基线结果的混淆矩阵可视化见图 1。

表 3　3 种模型 baseline 结果对比分析

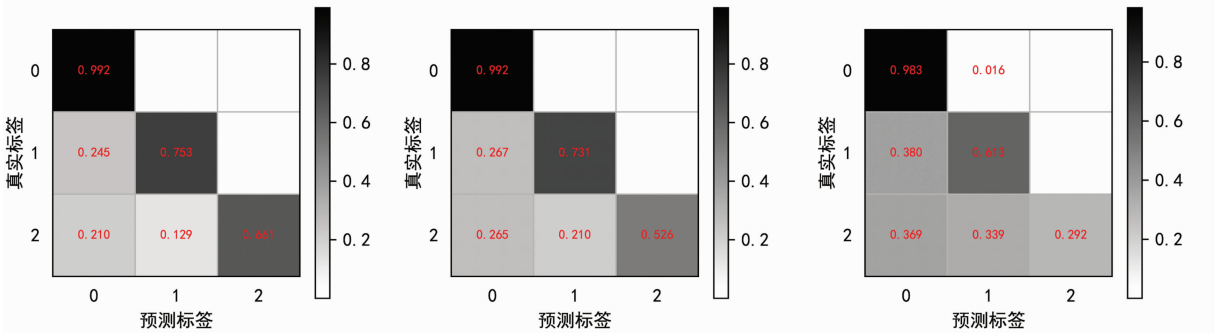
模型	类别	精确率（%）	召回率（%）	F_1 值	运行时间（s）
XGBoost	0	96.27	99.07	0.976 5	70.0
	1	90.35	75.33	0.821 6	
	2	97.28	66.05	0.786 8	
	加权平均值	95.55	95.67	0.954 7	
LightGBM	0	95.86	99.05	0.974 3	15.2
	1	89.01	73.08	0.802 6	
	2	97.92	52.61	0.684 5	
	加权平均值	95.05	95.19	94.900 0	
CatBoost	0	94.21	98.35	0.962 3	66.0
	1	79.89	61.30	0.693 7	
	2	79.80	29.15	0.427 0	
	加权平均值	92.23	92.78	0.921 4	

由混淆矩阵可以很清晰发现，3 种模型对类别 0（已脱贫）贫困人口识别效果非常好，均能达到 98% 以上的准确率。对类别 1（未脱贫）贫困人口识别稍差，XGBoost 与 LightGBM 可以达到 70% 以上，而 CatBoost 只有 61.3%。在对类别 2（返贫）的贫困人口识别上，XGBoost 最好，能够达到 66.1%，LightGBM 能够达到 52.6%，有一定的识别能力，CatBoost 分类效果较差，几乎是随机预测，不能够有效进行识别。

3.3　模型调优及分析

根据基线模型结果选择 XGBoost 和 LightGBM 等 2 个结果较为相近且效果较好的模型进行进一步调优比较。

（1）对 XGBoost 模型采用网格搜索（GridSearchCV）方法[13]以及五折交叉验证进行关键参数调优。最优参数见表 4。



a. XGBoost 混淆矩阵　b. LightGBM 混淆矩阵　c. GatBoost 混淆矩阵

图1　3 种模型基线测试结果的混淆矩阵对比

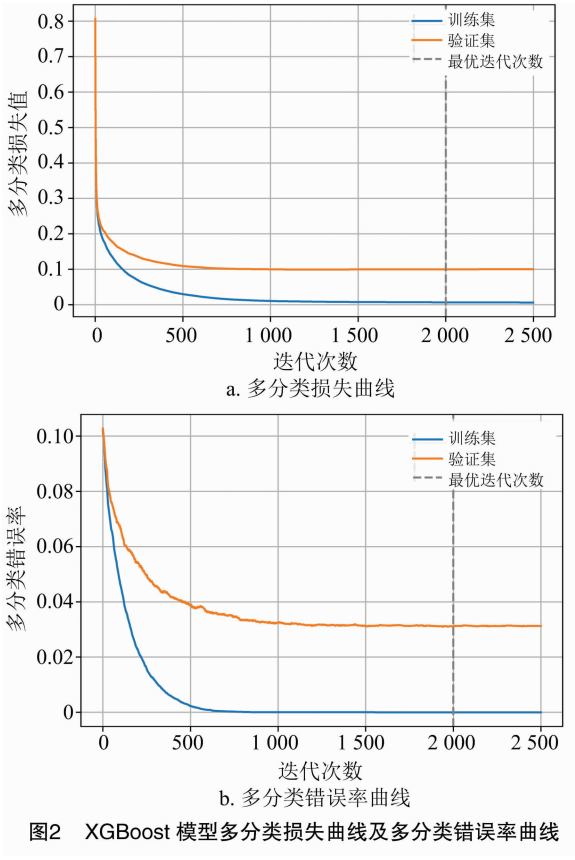
表 4　XGBoost 模型参数

参数	最优值	说明
学习率	0.35	控制每次迭代更新权重时的步长
总迭代次数	2 000	生成最大树的数量,即决策树个数
最大树深	6	树的深度越大,越容易过拟合
采样比例	0.95	训练每棵树时,使用数据占全部训练集的比例
惩罚系数	0.1	指定节点分裂时所需要的最小损失函数下降值
L1 正则化系数	0.0	模型各个参数的绝对值之和,控制模型防止过拟合
L2 正则化系数	0.2	模型各个参数的平方和的开方值,控制模型防止过拟合

XGBoost 模型在设置为表 4 中最优参数时,模型在测试集上的总体分类正确率达到 96.87%,相比较基线有 1.20% 的提升。模型训练的损失及错误率曲线见图 2。在迭代次数 2 000 次后,模型损失和错误率不再有明显下降,再增加迭代次数只会加大模型复杂度,增加模型过拟合概率。

(2) 对 LightGBM 模型采用网格搜索(GridSearchCV)方法[14]以及五折交叉验证进行关键参数调优。最优参数如表 5 所示。从表 5 可以看出,LightGBM 模型在设置中最优参数时,模型在测试集上的总体分类正确率达到 96.55%,相比较基线有 1.31% 的提升。模型训练的损失及错误率曲线见图 3。在迭代次数 2 200 次后,模型在验证集的损失有增加趋势,为过拟合产生的表现,不适宜再增加迭代次数。

XGBoost 模型与 LightGBM 模型经过调优后的各类别指标对比结果(表 6)显示,XGBoost 模型在各类别精确率以及召回率上均有微弱优势。混淆矩阵对比图见图 4,XGBoost 模型总体分类准确率比 LightGBM 模型高 0.32%,对于类别 0(已脱贫)和类别1(未脱贫)的分类准确率差距很小,只有 0.2% ~



a. 多分类损失曲线

b. 多分类错误率曲线

图2　XGBoost 模型多分类损失曲线及多分类错误率曲线

0.3%;对于类别 2(返贫)的分类准确率,XGBoost 模型比 LightGBM 模型高 3.7%,有较为明显的差距,但是其训练运行时间约为 LightGBM 模型的 4 倍。

2 个模型的特征重要性评估比较见图 5,XGBoost 和 LightGBM 等 2 个模型对特征重要性排序是完全一致的,仅仅是不同特征重要性值不同,排在前 5 的特征均为平均收入、住址、年龄、家庭劳动人口比率以及家庭人口数。根据特征重要性反映,在进行贫困类别判定时,更应该关注贫困人口收入、住址、家庭人口数以及健康医疗相关属性,着力加强这些方面的帮扶能够帮助贫困人口尽早脱

表 5　LightGBM 模型参数

参数	最优值	说明
学习率	0.2	控制每次迭代更新权重时的步长
总迭代次数	2 200	生成的最大树的数量，即决策树个数
最大叶子数	35	用叶子数量控制树的深度，值越大，越容易过拟合
叶子最小样本数	18	叶子节点可能具有最小记录数，设置较大值可以防止树生长过深，避免过拟合
采样比例	1	训练每棵树时，使用数据占全部训练集的比例
L1 正则化系数	0.1	控制模型防止过拟合
L2 正则化系数	0.1	控制模型防止过拟合

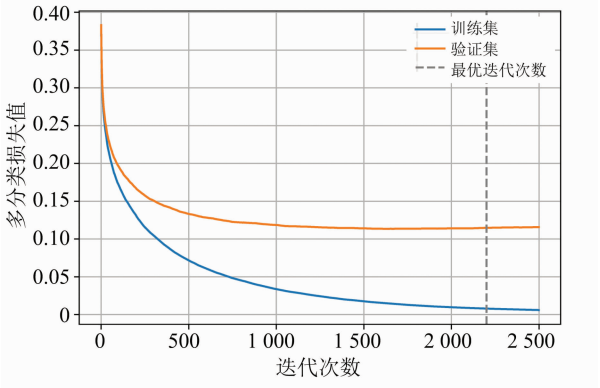


图3　LightGBM 模型多分类 logloss 曲线及多分类 error 曲线

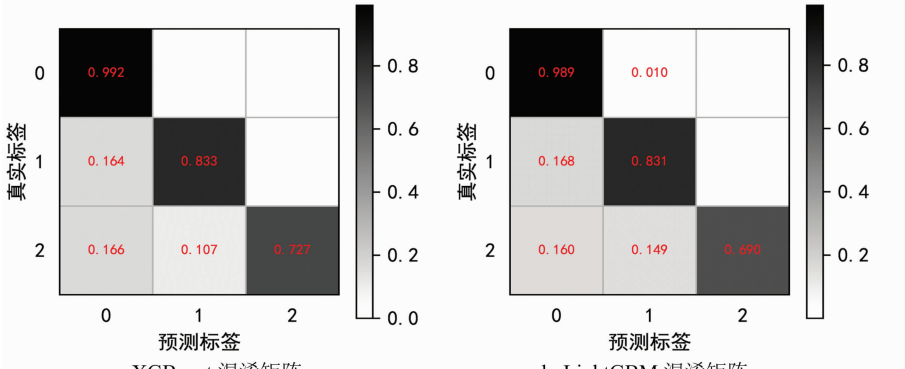
表 6　2 种模型调优后结果对比分析

模型	类别	精确率（%）	召回率（%）	F_1 值	运行时间（s）
XGBoost	0	97.43	99.21	0.983 1	74.0
	1	92.44	83.33	0.876 5	
	2	97.04	72.69	0.831 2	
	加权平均值	96.81	96.87	0.967 8	
LightGBM	0	97.38	98.92	0.981 5	18.9
	1	90.17	83.11	0.864 9	
	2	95.36	69.03	0.800 9	
	加权平均值	96.47	96.55	0.964 5	

贫。帮扶人员入户调查工作中，除填写一户一册帮扶手册以外还应及时上报更新帮扶对象的收入、家庭人口健康状况等信息。通过最新的贫困人口信息经由模型判断贫困人口最新的脱贫状态，以及追踪贫困人口贫困状态变化的最新影响因素。

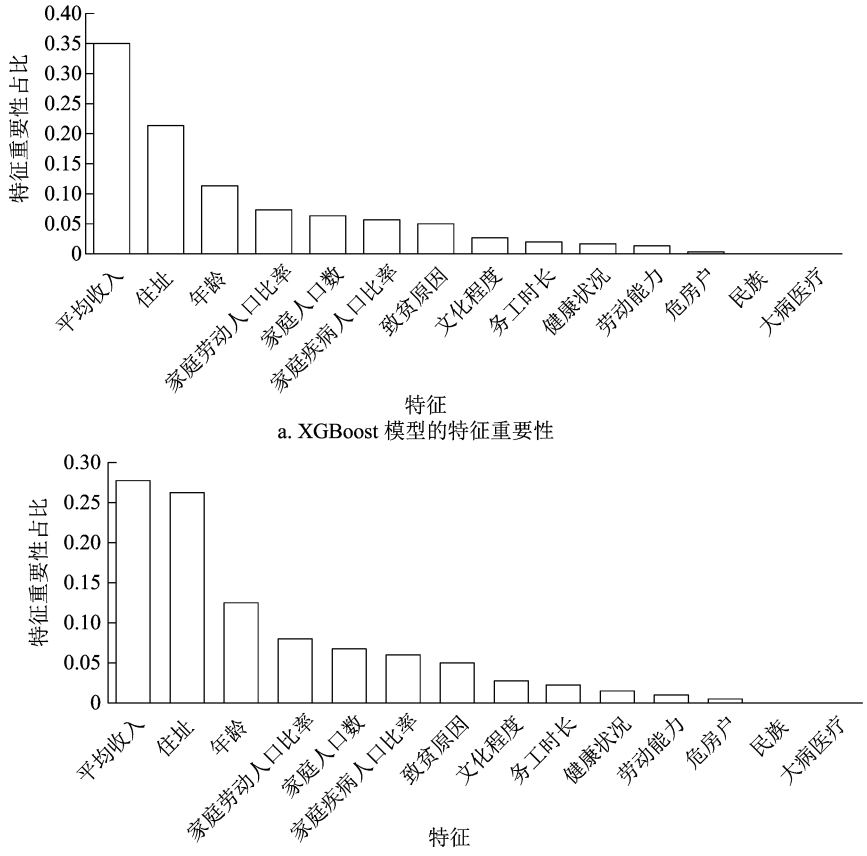
4　结论

本研究利用团队精准扶贫工作中积累的贫困户建档立卡数据，从中抽取 14 维特征，构建了基于集成学习的返贫人口识别模型，采用混淆矩阵、准确率以及 f_1 值等多指标对返贫人口识别模型进行了对比分析，基于 XGBoost 算法的返贫人口识别模型能够利用建档立卡数据对已脱贫、未脱贫及返贫 3 类人员分别达到 97.43%、92.44%、97.04% 的识别准确率，总体达到 96.81% 的准确率，能够较好识别出贫困人口贫困类别。通过构建基于集成学习算法的返贫人口识别模型，激活精准扶贫沉淀数据，为后脱贫时代的返贫动态监测预警工作提供实际案例支持，对我国由脱贫攻坚向全面推进乡村振



a. XGBoost 混淆矩阵　　b. LightGBM 混淆矩阵

图4　2 种模型调优后测试结果混淆矩阵对比



a. XGBoost 模型的特征重要性

b. LightGBM 模型的特征重要性

图5　2 种模型的特征重要性

兴平稳过渡有重要意义。本研究仍存在不足之处，如对贫困户数据采集维度较少，粒度较粗、数据类别存在不均衡等。在今后的防返贫工作中，要协调多方部门补充资产、政策补贴等数据，做到高时效、高精度防止返贫监测预警。

参考文献：

[1]李　洪，蒋龙志，何思好. 农村相对贫困识别体系与监测预警机制研究——来自四川省 X 县的数据[J]. 农村经济，2020，457(11)：69－78.

[2]范和生. 返贫预警机制构建探究[J]. 中国特色社会主义研究，2018，139(1)：57－63.

[3]罗　丽. 基于随机森林算法的贫困精准识别模型研究[J]. 华中农业大学学报(社会科学版)，2019，144(6)：21－29，160.

[4]杨　瑚. 返贫预警机制研究[D]. 兰州：兰州大学，2019.

[5] Ermon S. Combining satellite imagery and machine learning to predict poverty[J]. Science，2016(6301)：790－794.

[6]梁　骁，张　明，覃　琳. 一种基于机器学习识别贫困人口的数据分析方法研究[J]. 企业科技与发展，2017，427(5)：39－41.

[7]魏嫣娇，易叶青. 基于多源机器学习的脱贫方式智能推荐研究[J]. 信息与电脑(理论版)，2019，420(2)：37－39，44.

[8]张　浩. 提升农村地区精准扶贫效率的多维贫困识别方法[J]. 农村经济与科技，2020，31(6)：199－200.

[9]余　昕，汪早容. “后扶贫时代”返贫问题及对策[J]. 中国经贸导刊(中)，2021，992(1)：109－111.

[10]Chen T，Guestrin C. Xgboost：A scalable tree boosting system[C]// Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining(Association for Computing Machinery)，2016：785－794.

[11]Ke G，Meng Q，Finley T，et al. Lightgbm：A highly efficient gradient boosting decision tree[J]. Advances in Neural Information Processing Systems，2017，30：3146－3154.

[12]Dorogush A V，Ershov V，Gulin A. Catboost：Gradient boosting with categorical features support[J]. Arxiv E－prints，2018.

[13]岳　鹏，侯凌燕，杨大利，等. 基于 XGBoost 特征选择的疾病诊断 XLC－Stacking 方法[J]. 计算机工程与应用，2020，56(17)：136－141.

[14]陈维刚，张会林. 基于 RF－LightGBM 算法在风机叶片开裂故障预测中的应用[J]. 电子测量技术，2020，43(1)：162－168.