山西大学学报(自然科学版)47(1):155—165,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2023099

一种可解释的相对贫困识别与预警模型

史颖,丁天琪,祁晓博*,亓慧
(太原师范学院 计算机科学与技术学院,山西 晋中 030619)

摘 要:构建相对贫困人口识别体系和监测预警机制,增强监测预警机制的针对性和有效性,是相对贫困长效治理的必要条件。由于相对贫困识别的理论和算法研究相对较少,本文提出了一种可解释的相对贫困识别与预警模型,即IEWRP模型。该模型以2018—2020年中国家庭追踪调查问卷(CFPS2018)数据为研究对象,运用方差分析技术对原始数据集进行特征选择。然后通过Gradient Boosting构建IEWRP模型,并与决策树(Classification and Regression Tree,CART)、XGBoost、LightGBM等机器学习算法进行实验对比。最后,结合SHapley Additive exPlanation模型对影响相对贫困识别的相关特征的必要性和重要度进行了可解释性分析,识别出影响相对贫困识别的主要特征。实验结果显示,IEWRP模型的预测准确率、精确率、召回率、*F*1值、AUC(Area Under Curve)值分别为89.4%、90.6%、95.3%、92.9%、0.95,在准确率、精确率、*F*1值、AUC值四个方面分别提升了0.15%、0.28%、0.09%、0.23%。

关键词:Gradient Boosting模型;SHAP模型;相对贫困预测;特征分析
中图分类号:TP391 **文献标志码**:A **文章编号**:0253-2395(2024)01-0155-11

An Explainable Model for Identification and Early Warning of Relative Poverty

SHI Ying, DING Tianqi, QI Xiaobo*, QI Hui
(*College of Computer Science and Technology, Taiyuan Normal University, Jinzhong* 030619*, China*)

Abstract: The necessary conditions for long-term governance of relative poverty are to build a relative poor population identification system and a monitoring and early warning mechanism, as well as to enhance the pertinence and effectiveness of the monitoring and early warning mechanism. Due to the relatively limited theoretical and algorithmic research on relative poverty identification, in this paper, we propose an explainable model for identification and early warning of relative poverty, *i.e.* IEWRP model. This model takes the data of China Household Tracking Questionnaire (CFPS2018) from 2018 to 2020 as the research object, and uses technique of variance analysis to select features from the original data set. Then, the IEWRP model is constructed by gradient boosting, and compared with machine learning algorithms such as CART (Classification and Regression Tree), XGBoost and LightGBM. Finally, an interpretability analysis is conducted on the the necessity and importance of relevant features that affect relative poverty recognition with the shapley additive explanation model, identifying the main features that affect relative poverty recognition. Experimental results show that the prediction accuracy, precision, recall rate, *F*1 value and AUC (Area Under Curve) value of IEWRP model are 89.4%, 90.6%, 95.3%, 92.9% and 0.95, respectively. The accuracy, precision, *F*1 value and AUC value are increased by 0.15%, 0.28%, 0.09% and 0.23%, respectively.

Key words: gradient boosting model; shapley additive explanation model; relative poverty prediction; feature analysis

收稿日期:2023-04-06;**接受日期**:2023-05-18
基金项目:山西省哲学社会科学规划课题(2021YJ078);国家自然科学基金(62276161);山西省基础研究计划(自由探索)项目(20210302123334);山西省高等学校科技创新项目(2021L443)
作者简介:史颖(1990-),女,山西长治人,硕士,讲师,主要研究方向为机器学习,人工智能、生物信息等。E-mail:sy@tynu.edu.cn
* **通信作者**:祁晓博(QI Xiaobo),E-mail:xbqi@tynu.edu.cn
引文格式:史颖,丁天琪,祁晓博,等.一种可解释的相对贫困识别与预警模型[J].山西大学学报(自然科学版),2024,47(1):155-165. DOI:10.13451/j.sxu.ns.2023099

0 引言

2021年中国首次完成了消除绝对贫困的艰巨课题[1-3],扶贫事业的重心转变为解决相对贫困问题。相对贫困的精准识别和预警是反贫困工作的基础[4-5],如何巧妙运用机器学习算法进行数据处理,提高相对贫困预测精度,同时结合可解释性模型分析相对贫困的主要因素和特征,有利于相对贫困人群管理、识别、预警、帮扶工作的开展。

当前,相对贫困识别与预警研究主要围绕相对贫困识别的多维评价指标选择和相对贫困预测两个问题展开。汪三贵等[6]认为应根据人的基本需求来确定相对贫困标准,而王小林等[7]、檀学文[8]认为应建立多维贫困标准。典型的相对贫困多维指标选择仅在理论层面进行了分析[9-10],缺乏数据及实验支撑。现有的相对贫困预测方法,如基于随机森林[11]的机器学习方法虽然在中国家庭追踪调查(CFPS)数据上建立了预测模型,但未与其他模型对比,且仅以精准率作为评价指标,未考虑该模型的泛化能力。

逻辑回归(Logistic Regression,LR)、K近邻法(K-Nearest Neighbor,KNN)、支持向量机(Support Vector Machine,SVM)、随机森林(Random Forest,RF)、决策树(Classification and Regression Tree,CART)等常见的机器学习算法都可以用于分类和回归问题,但是对于噪声和异常值比较敏感,部分算法在处理高维、大规模数据集时表现不佳,容易过拟合。Gradient Boosting算法善于处理高维稀疏数据、非线性关系、异常值和噪声数据,适用于分类和回归问题。由于相对贫困识别预测问题通常具有高度非线性和复杂性,该算法可以更好地捕捉数据之间的非线性关系和复杂交互作用,相对于其他机器学习算法具有更高的准确性和鲁棒性。因此,基于Gradient Boosting构建一种可解释的IEWRP模型(An Explainable Model for Identification and Early Warning of Relative Poverty,简称IEWRP模型)具有很大的优势。

针对相对贫困识别和预测实证层面研究不充分、模型可解释性不强的问题,本文基于2018—2020年中国家庭追踪调查问卷(CFPS2018)数据集,在对数据进行处理后,进行特征关联分析,选择出较为合适的训练特征,并进行了模型训练、超参数优化,接着进行了模型性能分析和模型解释,最后在这些工作的基础上建立了具有良好性能的IEWRP模型。本文工作主要集中在以下3个方面:

1)运用单因素方差分析法及箱线图进行特征分析和选择,排除冗余特征,提高模型的预测性能。

2)基于Gradient Boosting算法建立了IEWRP模型,并通过网格搜索技术对模型超参数进行调优,进一步提高模型性能表现。对比实验证实了本文模型在多个预测指标上具有优越性。同时,通过对比其他模型的学习曲线,证明了本文模型具有较好的泛化能力。

3)在此基础上,以SHAP(SHapley Additive exPlanation)模型为基础解析影响相对贫困的主要特征,以提高模型的可解释性。

1 相关技术

1.1 特征关联分析

特征关联分析(Feature correlation analysis)是一种用于分析特征之间相关性的方法。特征关联分析通常用于数据预处理和特征选择,可以帮助了解特征之间的相互影响关系,排除冗余特征,选择与目标变量相关性高的特征,提高模型的预测性能。常用的特征关联分析方法包括:相关系数分析、热力图可视化、方差分析[12-13]。

方差分析法是一种用于比较两个或多个样本平均值是否有显著差异的统计方法。其基本思想是比较样本内部的差异与样本之间的差异,以判断不同组之间的差异是否显著。运用单因素方差分析进行特征选择是为了测试某一特征变量在不同水平程度下目标变量是否会产生显著差异和变动,从而实现有效的特征选择。通过方差p值进行特征关联分析的步骤如下:

第一步对于每个特征,根据目标变量的取值将数据分成若干组。第二步对于每个分组,计算该分组内特征的方差,以及所有分组内特征的均值和方差。第三步根据均值和方差计算

组间方差和组内方差,并计算F统计量,即不同组之间的方差占整体方差的比例。第四步根据F统计量计算p值。

1.2 梯度提升

1.2.1 决策树

决策树是一种基于树结构的分类和回归模型,通过对数据进行递归的分割,最终得到一系列叶节点,每个叶节点代表了一个分类或回归的结果。分类决策树是一种基于树状结构进行决策的机器学习算法,用于处理分类问题。

分类决策树将数据集按照某种特定方式分成多个子集,每个子集对应一个判定条件,形成一棵树状结构,从根节点到叶节点表示分类决策的过程[14-16]。通常采用自顶向下的贪心算法,每次选择一个最佳的特征作为节点进行划分,使得每个子集的纯度增加,同时保证划分后的子集互斥且包含全部样本。这个过程一直持续到某个终止条件满足为止,例如达到最大深度、无法继续划分、样本量太小等。

决策树的优点在于易于理解和解释,能够处理离散型和连续型数据,对缺失值不敏感,且能够自动进行特征选择。另外,决策树算法也可以用于集成学习中的随机森林和梯度提升决策树等算法中[17]。

1.2.2 梯度提升算法

Gradient Boosting是一种基于决策树的集成学习方法,可以用于构建分类预测模型。梯度提升算法是一个迭代的优化过程,它通过不断地训练模型来最小化误差。每个模型都是在之前模型的基础上进行训练,从而逐步提高模型的准确度。由于每个模型都只是一个简单的模型,梯度提升算法在实践中通常会使用多个模型进行组合,以进一步提高模型的准确度。

1.2.3 梯度提升决策树

梯度提升决策树(Gradient Boosting Decision Tree,GBDT)是一种使用集成学习思想的算法,它通过迭代地训练一组基础决策树模型并将它们进行组合,以提高模型的准确率和泛化能力[18]。GBDT的每个模型都是在之前模型的基础上进行训练,从而逐步提高模型的准确度。在每一次迭代中,GBDT会根据前一次迭代的结果对数据进行加权,从而使每次迭代都着重于改进先前的错误,经多次迭代后使得残差趋近于0。最终,GBDT会将多个模型的预测结果进行加权组合,得到一个更为准确的预测结果[19]。GBDT算法的一般流程如下,首先使用训练集训练模型,通过迭代的方式训练多个决策树模型,并逐步优化模型的参数和结构,使得每个决策树都可以更好地拟合训练集数据,并尽可能地减少误差[20]。

1.3 SHAP模型

SHAP是一种用于解释机器学习模型预测的框架,可以提供单个样本的特征重要性分析和模型预测的可解释性。SHAP框架的核心思想是将机器学习模型的预测结果拆解为各个特征的贡献值,从而提供每个特征对模型预测结果的贡献程度和解释[21-22]。

通过SHAP框架提供的可视化工具,我们可以更好地理解模型预测结果的原因和特征重要性,从而提高模型的可解释性和可靠性[23-24]。计算SHAP值不仅可以将数据可视化来展示模型预测的结果,还能表现出特征对预测结果影响的正负性,即可以通过特征对最终预测值的贡献值是否大于0来判断该特征是否提升了预测值,若大于0则提升了预测值,若小于0则表明该特征降低贡献。

2 可解释的相对贫困风险预测及特征分析

IEWRP模型对相对贫困风险进行预测并对其特征进行分析,该模型构建流程如图1所示,主要包括将处理后的数据进行特征关联分析、训练模型并进行超参数优化、模型性能比较和模型解释分析等模块。

具体包括以下几个主要步骤。(1)特征关联分析:对原始数据进行预处理,接着进行特征转换等工作,再运用箱线图分析、简单相关性分析、方差分析法进行特征关联分析,选取合适的训练特征,降低模型训练时间。(2)构建相对贫困预测模型:通过Gradient Boosting构建相对贫困预测模型,并进行10折交叉验证、超参数优化等工作获得最优的模型架构。(3)基于SHAP的模型解释:结合SHAP模型分析不同类别中影响相对贫困识别的主要特征,为将模型应用到实际相对贫困治理方案打下基础。

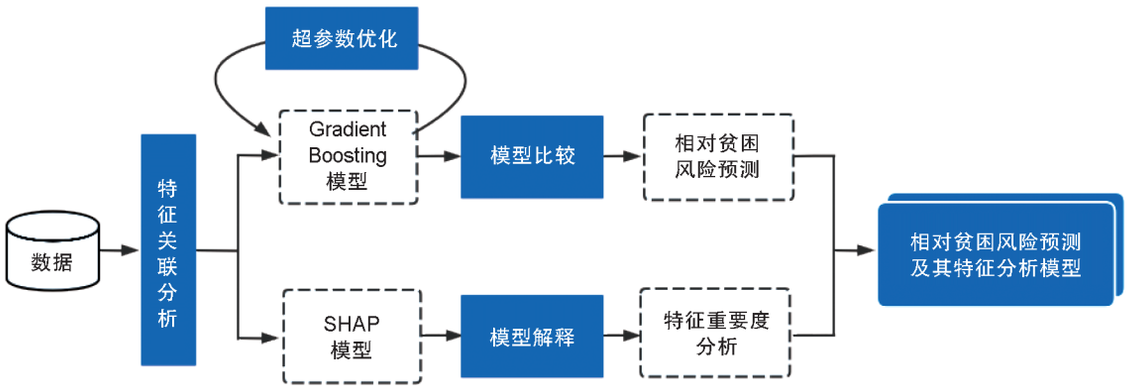


图1 IEWRP模型流程图

Fig. 1 Flow chart of IEWRP model

2.1 特征关联分析

在对原始贫困调查问卷数据进行分析时,由于相对贫困可能受到多种因素影响,例如教育、健康、生活水平、收入等,但是上述多种因素中部分特征对于年收入和判断是否相对贫困影响较小,因此可以选取具有可靠性、关联性和代表性的数据,以减少数据处理工作并排除无关特征的干扰。

以“成人健康”特征为例,原始数据中将成人健康程度划分为5个等级,不同健康等级可能影响家庭的相对贫困结果。为研究不同特征取值对相对贫困的影响是否显著,本文通过方差p值进行特征关联分析。设样本数为n,因素有k个水平,每个水平的均值分别用$\mu_1,\mu_2,\cdots,\mu_k$表示。通过方差p值进行特征关联分析的步骤如下:

第一步:提出如下假设。

$H_0:\mu_1=\mu_2=\cdots=\mu_k$,$H_1:\mu_1,\mu_2,\cdots,\mu_k$不全相等。

第二步:构造检验的统计量。

(1)计算总误差平方和S_T、组间平方和S_A、组内平方和S_E

$$S_T=\sum_{i=1}^{k}\sum_{j=1}^{n_{ij}}\left(x_{ij}-\overline{\overline{x}}\right)^2, \tag{1}$$

$$S_A=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\overline{x_i}-\overline{\overline{x}}\right)^2, \tag{2}$$

$$S_E=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{x_i}\right)^2, \tag{3}$$

其中,$\overline{\overline{x}}$为全部观察值的总均值,n_i为第i个总体的样本观察值个数,$\overline{x_i}$为水平的均值,x_{ij}为第i个总体的第j个观察值。

(2)计算组间均方$\overline{S_A}$、组内均方$\overline{S_E}$

$$\overline{S_A}=\frac{S_A}{k-1}, \tag{4}$$

$$\overline{S_E}=\frac{S_E}{n-k}。 \tag{5}$$

(3)计算统计量F

$$F=\frac{\overline{S_A}}{\overline{S_E}}。 \tag{6}$$

第三步:统计决策。

若$F>F_\alpha$,则拒绝原假设H_0,表明均值之间的差异是显著的,所检验的因素对观察值有显著影响。若$F<F_\alpha$,则不拒绝原假设H_0,无证据表明所检验的因素对观察值有显著影响。

p值是指在零假设成立的情况下,观察到当前的F值或更极端情况下,所得到的结果的概率。如果p值小于一定的显著性水平(通常为0.05),那么就可以拒绝原假设,认为这个特征与目标变量之间存在显著差异,即两者之间存在关联。否则认为不存在关联。

2.2 Gradient Boosting算法

基于Gradient Boosting算法构建相对贫困识别预测模型,设全国相对贫困数据集为X(包括年收入、家庭人数、成人教育、成人健康等特征),Y为人均年收入(单位为元),将数据集划分为训练集与测试集数据,给定n个样本(表示n个个体)、m个特征的训练集数据为

$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_i,y_i),\cdots,(x_n,y_n)\}$,

其中$x_i=(x^1,x^2,\cdots,x^m)$。将训练集数据输入Gradient Boosting算法模型进行训练,对测试集数据进行预测,并根据算法评价指标评估模型

的性能表现。针对一般的损失函数求解优化过程较为复杂的问题,Friedman提出了梯度提升算法(Gradient Boosting)。该算法的核心就是利用基于当前模型下损失函数的负梯度的取值来近似拟合残差,形成一棵回归树。梯度提升算法如下所示:

算法:梯度提升算法

输入:训练集数据

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\},$$

$x_i\in X\subseteq R^n, y_i\in Y\subseteq R$,损失函数$L(y,f(x))$;

输出:回归树$\hat{f}(x)$。

第一步:初始化

$$f_0(x)=\arg\min_c\sum_{i=1}^{N}L(y_i,c)。\tag{7}$$

这一步通过估计使损失函数极小化的常数值,得到一个根结点的树。

第二步:对$m=1,2,\cdots,M$进行迭代

(1)对$i=1,2,\cdots,N$计算负梯度值:

$$r_{mi}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}。\tag{8}$$

(2)对r_{mi}拟合一棵回归树,得到第m棵树的叶节点域R_{mi},$j=1,2,\cdots,J$。

(3)对$j=1,2,\cdots,J$计算:

$$c_{mj}=\arg\min_c\sum_{x_i\in R_{mj}}L(y_i,f_{m-1}(x_i)+c),\tag{9}$$

更新回归树

$$f_m(x)=f_{m-1}(x)+\sum_{j=1}^{J}c_{mj}I(x\in R_{mj})。\tag{10}$$

第二步首先将所得的结果r_{mi}作为残差的估计。再对回归树叶结点区域进行估计,对残差进行拟合。接着通过线性搜索使损失函数极小化,估计叶结点区域的值。最后更新回归树。

第三步:得到回归树

$$\hat{f}(x)=f_M(x)=\sum_{m=1}^{M}\sum_{j=1}^{J}c_{mj}I(x\in R_{mj})。\tag{11}$$

2.3 模型可解释性

基于Gradient Boosting算法进行训练后得到的预测模型在预测精度方面比较可观,在模型的可解释性方面效果不佳。因此,本文采用SHAP值来对模型中影响相对贫困的特征进行解释分析,用来增强模型的可解释性。模型预测值为所有样本的预测均值及各特征的SHAP值之和,公式如下:

$$y=f_{base}+f_1+\cdots f_i+\cdots+f_M,\tag{12}$$

其中,y为该模型的预测值,f_i为特征i对应的SHAP值,f_{base}为所有样本预测值的均值。通过计算SHAP值来增强可解释性的最大优点是能准确地反映出样本中每个特征对模型的影响,并能准确地判断出这些特征对模型的影响程度以及影响的正负性。

3 实验结果及分析

3.1 特征关联分析及选择

本实验的数据集2018—2020年中国家庭追踪调查问卷(CFPS2018)数据,包括个体的住房、年收入、家庭人数、年龄、医保及使用炊用燃料情况等多项信息内容。影响识别预测是否相对贫困的属性信息包含收入、教育、健康、生活水平4个类别,细分为成人教育、儿童教育、成人健康、儿童健康、医疗保险、炊用燃料、安全饮用水、住房、资产、年收入10个特征属性,依次用x1—x10表示。人均年收入有两种表示,一种是实数型的实际年收入(y0),另一种是类别型年收入(y1),低于相对贫困阈值的取1,其余取0。数据集所含部分特征如表1所列。

人均年收入的范围在0~800 000元之间,人均年收入的平均值为24 273元,家户人均可支配收入分布图如图2(a)所示,数据分布集中在左侧,为了使模型在处理数据时效果更佳,本文通过将人均年收入取对数使目标值正态化,变换后家户人均可支配收入分布如图2(b)所示。

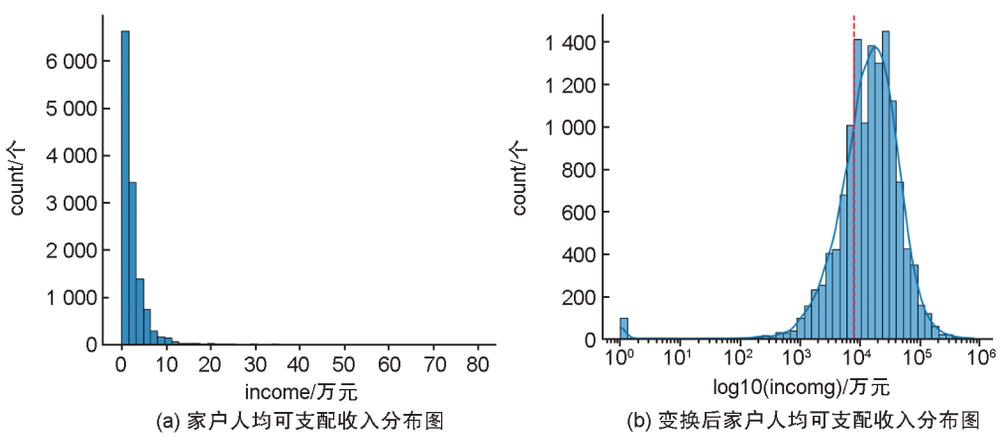
图3展示了各个特征与人均年收入的相关性箱线图。箱线图中箱子中间的红色横线是数据的中位数,代表了样本数据的平均水平。箱子的上下范围在一定程度上反映了数据的波动程度。由图3可以看出,x1和x5取值不同时,y0值相差不大;而对于其他特征取值不同时,y0值的箱线图差别较大,说明x1和x5特征对收入无明显影响,即除成人教育和医疗保险的其他特征对收入有明显影响。

p值用于检验特征与变量之间相关性,如果

表1 相对贫困数据描述性统计

Table 1 Descriptive statistics of relative poverty data

变量	意义	count	mean	std	min	0.25	0.50	0.75	max
x1	成人教育	13 056	1.00	0.04	0.0	1.0	1.0	1.0	1.0
x2	儿童教育	13 056	0.99	0.12	0.0	1.0	1.0	1.0	1.0
x3	成人健康	13 056	2.83	1.22	1.0	2.0	3.0	3.0	5.0
x4	儿童健康	13 056	3.51	1.01	1.0	4.0	4.0	4.0	4.0
x5	医疗保险	13 056	0.89	0.32	0.0	1.0	1.0	1.0	1.0
x6	炊用燃料	13 056	3.42	1.72	1.0	2.0	4.0	4.0	6.0
x7	安全饮用水	13 056	2.74	0.59	1.0	3.0	3.0	3.0	4.0
x8	住房	13 056	2.59	0.79	1.0	3.0	3.0	3.0	3.0
x9	资产	13 056	47 613.81	117 127.11	0.0	3.20×10^3	1.00×10^4	4.38×10^4	5.00×10^6
x10	年收入	13 056	64 754.79	80 149.61	0.0	2.00×10^4	4.79×10^4	8.00×10^4	2.00×10^6
y0	实数型实际年收入	13 056	24 273.32	32 475.22	0.0	7.50×10^3	1.50×10^4	3.00×10^4	8.00×10^5
y1	类别型实际年收入	13 056	0.72	0.45	0.0	0.0	1.0	1.0	1.0



(a) 家户人均可支配收入分布图 (b) 变换后家户人均可支配收入分布图

图2 居民人均可支配收入分布图

Fig. 2 Distribution map of Household disposable income per capita

p值小于一定的显著性水平(通常为0.05),那么就可以拒绝原假设,认为这个特征与目标变量之间存在显著差异,即两者之间存在关联。计算所有特征的p值进行特征选择,特征的p值如表2所示。

通过特征相关性箱线图及计算特征的p值进行特征选择。众多特征中x1和x5对应的p值均大于0.05,说明x1和x5在不同取值下,y0值不存在显著差异,即成人教育和医疗保险没有呈现出显著性,因此将成人教育和医疗保险两个特征剔除。其中,对x9和y0进行相关性检验,p=0.000,相关系数为0.33,表明存在显著相关性。同理,对x10和y0进行相关性检验,p=0.000,相关系数为0.77,存在显著相关性。故所选特征为x2,x3,x4,x6,x7,x8,x9,x10,即儿童教育、成人健康、儿童健康、炊用燃料、安全饮用水、住房、资产、年收入。

3.2 相对贫困识别预测

3.2.1 模型评价指标

本文基于x2,x3,x4,x6,x7,x8,x9,x10和y1建立分类模型,主要采用如下分类性能指标来评估模型:

(1)准确率(accuracy):表示预测分类正确的样本占样本总量的百分比,其计算公式如式(13):

$$accuracy=\frac{TP+TN}{TP+TN+FP+FN}。\tag{13}$$

(2)精准率(precision):表示正样本中预测分类正确的样本所占被预测为正样本的百分比,其计算公式如式(14):

$$precision=\frac{TP}{TP+FP}。\tag{14}$$

(3)召回率(recall):预测分类正确的正样

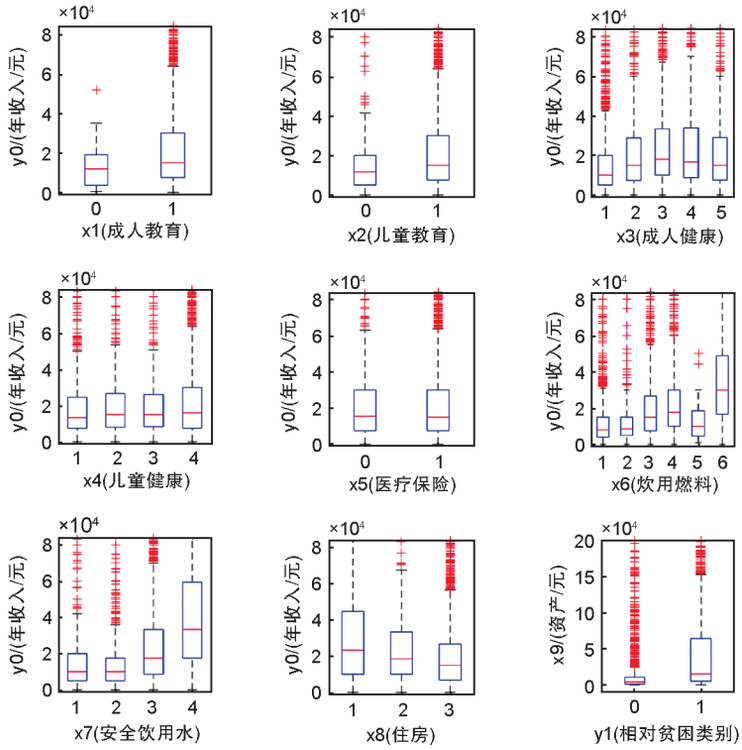


图3 部分特征与人均年收入的相关性箱线图

Fig. 3 Box plot of correlation between partial characteristics and annual income per capita

表2 相对贫困数据各特征的p值

Table 2 p-values for each feature of relative poverty data

变量名称	变量解释	p值
x1	成人教育	0.137 2
x2	儿童教育	0.000 6
x3	成人健康	0.000 0
x4	儿童健康	0.000 0
x5	医疗保险	0.727 5
x6	炊用燃料	0.000 0
x7	安全饮用水	0.000 0
x8	住房	0.000 0
x9	资产	0.000 0
x10	年收入	0.000 0

本所占真实正样本的百分比,其计算公式如式(15):

$$recall=\frac{TP}{TP+FN}。\tag{15}$$

(4)$F1$值:综合评估精准率和召回率,当$F1$=1时,模型输出结果最优,反之,$F1$=0时,模型输出结果最差。其计算公式如式(16):

$$F1=\frac{2\times precision\times recall}{precision\times recall}\times 100\%。\tag{16}$$

(5)AUC值(Area Under Curve):受试者工作特征曲线(Receiver Operating Characteristic Curve,ROC)与横轴之间的面积和。AUC值越大,则模型精确度越好。

在式(13)—(16)中:TP表示正样本中预测分类正确的数目,FN是正样本中预测分类错误的数目,FP为负样本中预测分类错误的数目,TN为负样本中预测分类正确的数目。分类器在测试数据集上预测的结果可以通过表3所示的混淆矩阵来表示。

表3 相对贫困分类结果混淆矩阵

Table 3 Confounding matrix of relative poverty classification results

贫困状态	预测相对贫困	预测不相对贫困
实际相对贫困	*TP*	*FN*
实际不相对贫困	*FP*	*TN*

3.2.2 模型参数优化

Gradient Boosting有4个核心参数。通过调整设定合理的参数,可以提高模型的性能。本文运用网格搜索方法,同时结合经验判断和遍历实验来确定最佳参数,如表4所示。

3.2.3 模型对比

本节将Gradient Boosting与已有的LR(逻辑回归)、KNN、SVM、RF(随机森林)、CART、

表4 Gradient Boosting最优参数

Table 4 Gradient Boosting optimal parameters

参数名	默认值	参数优化值	参数含义
learning_rate	0.01	0.03	学习率
n_estimators	50	200	迭代次数
max_depth	3	5	决策树的最大深度
Subsample	1	0.5	随机抽样的时候抽取的样本比例

XGBoost、LightGBM 7种模型进行对比。实验中采用了10折交叉验证方法,以增强模型之间对比的公平性及可信度。以准确率的值作为评价指标绘制箱线图,如图4所示。由箱线图中的分布情况可以看出,Gradient Boosting模型的预测精度比已有的7种模型要高。

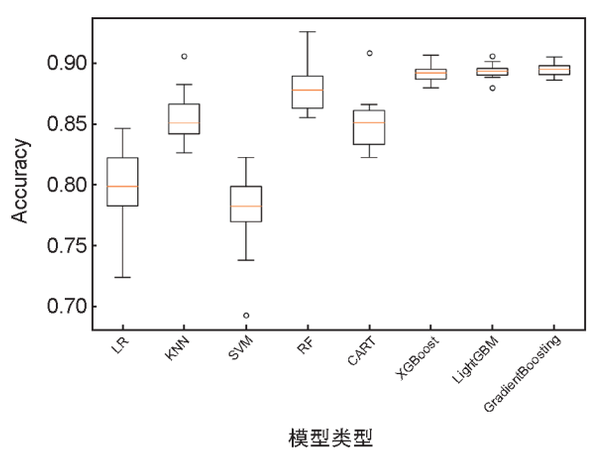


图4 各模型准确率比较

Fig. 4 Comparison of accuracy of each model

表5比较了不同模型在各个指标上的性能。由表可知,Gradient Boosting的预测准确率为89.4%,精确率为90.6%,召回率95.3%,*F*1值为92.9%,AUC值为0.95。进行参数优化后的PGradient Boosting在精确率、召回率、*F*1值和AUC值四个指标上都有了不同程度的提高。XGboost、LightGBM、Gradient Boosting在预测准确率上表现明显优于其他机器学习算法,预测准确率均达到了89%以上,体现了集成算法的优越性。在各项性能上,Gradient Boosting与LightGBM算法最为接近,但总体来看Gradient Boosting性能更优,提高了约0.1%~0.3%。原因在于Gradient Boosting以决策树为基学习器,通过抑制基学习器的复杂度,缓解基学习器的过拟合风险,提高模型泛化能力。此外,采用决策树作为弱分类器使得GBDT模型具有较好的解释性和鲁棒性,能够自动发现特征间的高阶关系,并且也不需要对数据进行特殊的预处理,如归一化。Gradient Boosting模型参数优化后性能提高约0.2%~1.0%。综合来看,本文提出的预测模型对于相对贫困的识别和预测能力表现良好。

表5 各模型性能对比(%)

Table 5 Performance comparison among models (%)

模型	accuracy	precision	recall	*F*1	AUC
LR	79.69	81.19	93.78	86.99	87.04
KNN	85.67	88.38	92.32	90.30	88.92
SVM	77.58	80.03	92.14	85.58	87.54
RF	87.95	90.65	92.92	91.76	93.67
CART	85.09	90.15	89.12	89.62	82.43
XGBoost	89.17	90.55	94.93	92.69	95.10
LightGBM	89.33	90.36	95.43	92.82	95.35
Gradient Boosting	89.48	90.64	95.31	92.91	95.58
PGradient Boosting	**89.63**	**89.95**	**96.44**	**93.08**	**95.58**

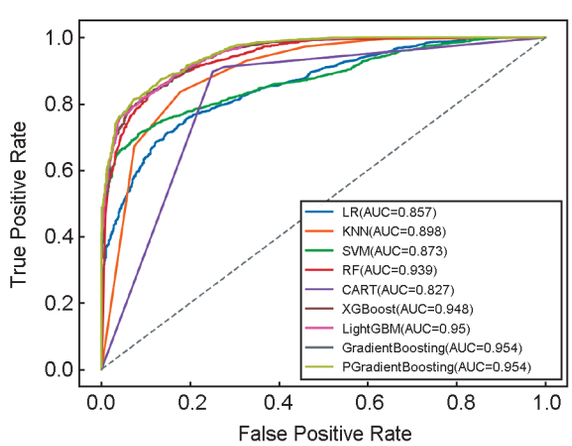


图5 各算法ROC曲线对比

Fig. 5 ROC curve comparison among algorithms

图5展示了不同模型在相对贫困识别中的ROC曲线。由图可见,Gradient Boosting的ROC曲线位于最右上方,表明此方法的预测效果最佳,与表5中的AUC结果一致。

衡量模型的好坏只考虑在训练集上的表现远远不够,还需进行交叉验证,验证其泛化能力如何。一个好的模型不仅要在训练集上表现优秀,还要对其他新的样本有一定的适应能力,达到较好的效果。本文通过学习曲线来分析模型的收敛情况,Gradient Boosting与其他模型的学习曲线对比情况如图6所示,其中,其他模型包括LR、SVM、LightGBM。

由图6可以看出,这4种算法在总体拟合趋

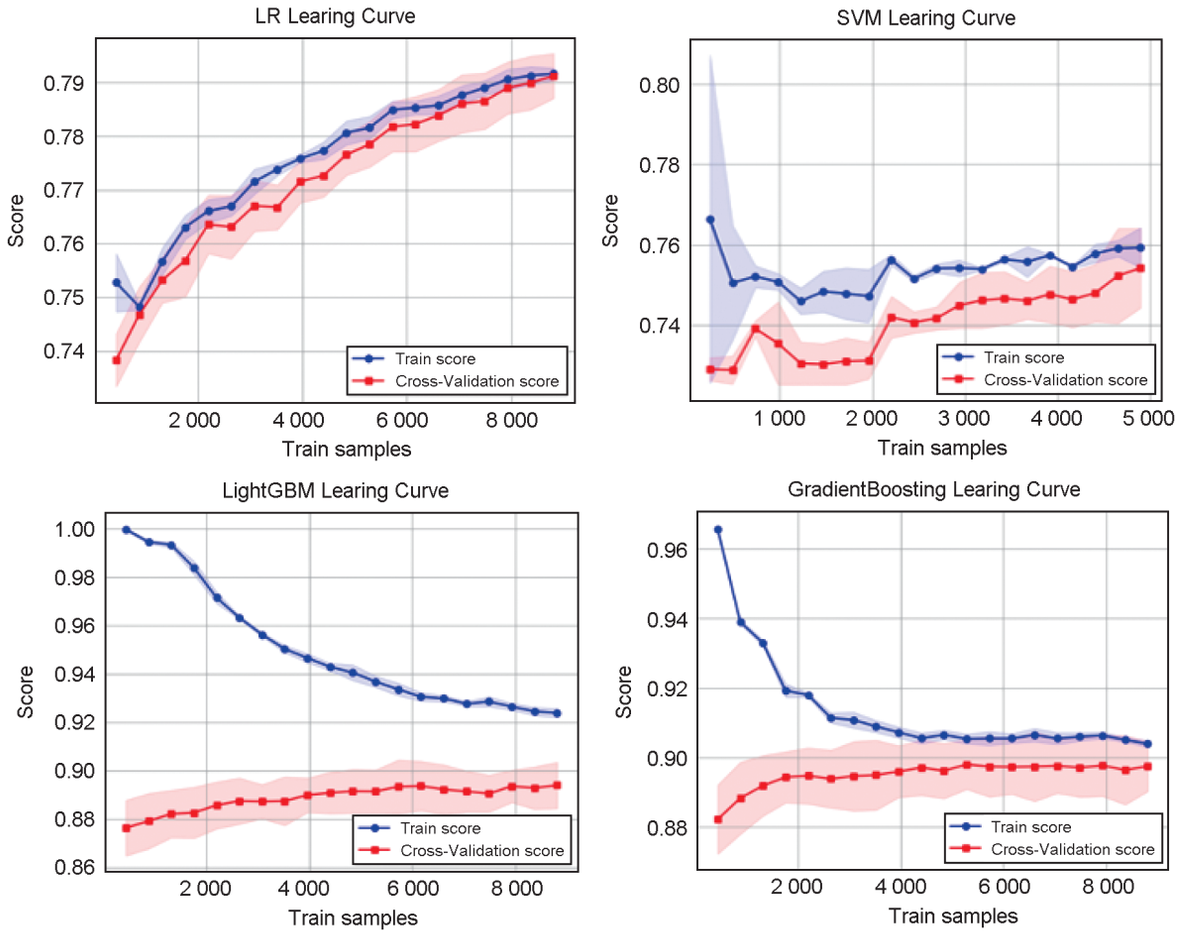


图6 各算法学习曲线对比

Fig. 6 Comparison among the learning curves of algorithms

势上呈现效果不同,LR算法随着样本数量增大训练得分也不断增大。而SVM、LightGBM、Gradient Boosting这三种模型在训练样本较少时训练得分非常高,测试得分非常低;当增加样本数量时,训练得分会有所下降,测试得分逐渐增加。此外,LightGBM算法和Gradient Boosting算法在任何训练样本数量下,训练得分和拟合效果都高于SVM和LR。总之,从整体拟合趋势来看,Gradient Boosting算法随着训练样本数量的增加,训练得分和测试得分都更加趋于稳定,拟合效果也较为优异。

3.3 基于SHAP的模型解释分析

图7显示了SHAP条形图,按每个特征的SHAP值进行排序,体现全局特征的重要性。从图7可以看出,年收入、儿童健康、炊用燃料等特征对相对贫困的分类影响较大,其中年收入特征的差异对模型的影响最为显著。

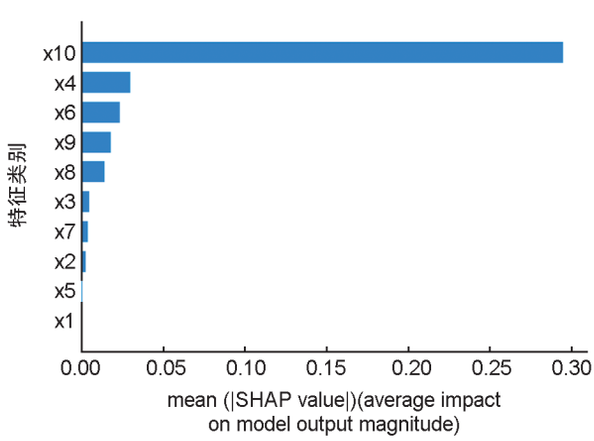


图7 影响相对贫困特征的重要性分析

Fig. 7 An analysis of the importance of influencing the characteristics of relative poverty

图8显示了SHAP摘要图,该图根据特征对影响认定是否相对贫困的特征重要性进行排序。从图8可以看到:年收入、儿童健康、炊用燃料这些特征对是否是相对贫困的预测都具有显著的影响。随着这些特征的值减小,认定为

相对贫困的风险就越小,因此这些特征对认定为相对贫困都具有负面影响。其中,年收入对影响认定是否相对贫困最为重要,低年收入对预测是负向影响,高年收入对预测是正向影响,即年收入越高的家庭情况越好,成为相对贫困的风险越小。而资产这个特征对预测具有正面影响,资产越少,认定为相对贫困的风险越大。

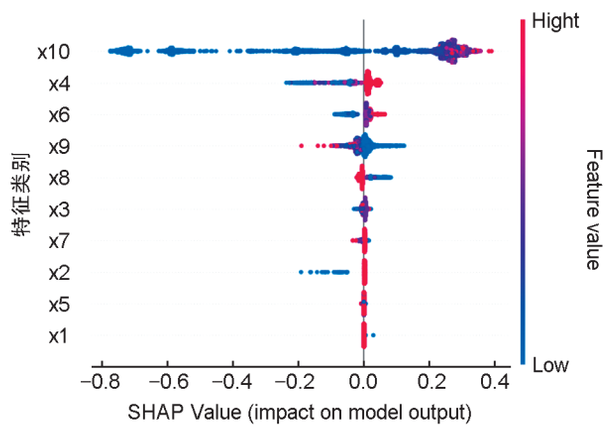


图8 SHAP特征分析摘要

Fig. 8 Summary of SHAP feature analysis

4 结论与展望

综合考虑相对贫困识别和预测模型实证层面不完备和模型可解释性较弱的问题,本文提出IEWRP模型。利用机器学习算法提高相对贫识别预测模型的性能,融合模型的可解释性识别影响相对贫困的特征。首先对原始数据集进行特征选择后将数据作为训练模型的输入,训练后的模型通过网格搜索方法寻找最优参数。通过与LR、KNN、SVM、RF、CART、XGBoost、LightGBM这7种机器学习模型的对比实验,证明了IEWRP模型的有效性及优异的拟合能力。最后利用SHAP模型进行特征解释分析,对特征重要性进行排序,识别出对相对贫困识别预测影响较大的几个关键特征,分别是年收入、儿童健康、炊用燃料、资产。针对识别出的相对贫困人群进行预警,根据影响较大的特征和人群进行精准扶贫,对于进一步开展相对贫困的治理工作具有重要的现实意义。

构建相对贫困风险预测模型并进行特征解释分析,实现相对贫困人群的识别和预警,对辅助相对贫困治理有极大帮助。下一步工作考虑结合各省份不同地区的具体需求对特征分析、超参数优化等内容进行优化,进一步提高相对贫困预测模型的效果,并将模型推广到实际场景,实现更为精准的相对贫困人群预警,发挥更大的应用价值。

参考文献:

[1] 习近平. 在全国脱贫攻坚总结表彰大会上的讲话[N]. 人民日报, 2021-02-26(2). DOI: 10.28655/n. cnki.nrmrb.2021.001951.

[2] 顾仲阳. 成就彪炳史册 奋斗开创未来[N]. 人民日报, 2022-03-14(7). DOI: 10.28655/n.cnki.nrmrb.2022.002547.

[3] 樊增增, 邹薇. 从脱贫攻坚走向共同富裕: 中国相对贫困的动态识别与贫困变化的量化分解[J]. 中国工业经济, 2021(10): 59-77. DOI: 10.19581/j.cnki.ciejournal.2021.10.010.
FAN Z Z, ZOU W. From Anti-poverty Campaign to Common Prosperity: Dynamic Identification of Relative Poverty and Quantitative Decomposition of Poverty Changes in China[J]. *China Ind Econ*, 2021(10): 59-77. DOI: 10.19581/j.cnki.ciejournal.2021.10.010.

[4] 郑继承. 构建相对贫困治理长效机制的政治经济学研究[J]. 经济学家, 2020(5): 91-98. DOI: 10.16158/j.cnki.51-1312/f.2020.05.010.
ZHENG J C. A Political Economics Study on Establishing a Long-term Mechanism for Relative Poverty Governance [J]. *Economist*, 2020(5): 91-98. DOI: 10.16158/j.cnki.51-1312/f.2020.05.010.

[5] 胡联, 杨成喻, 姚绍群. 数字经济对相对贫困的影响机制分析与政策启示[J]. 西安财经大学学报, 2023, **36**(3): 38-50. DOI: 10.19331/j.cnki.jxufe.2023.03.002.
HU L, YANG C Y, YAO S Q. The Impact of Digital Economy on Relative Poyerty: Mechanism Analysis and Policy Implications[J]. *J Xi'an Univ Financ Econ*. 2023, **36**(3): 38-50. DOI:10.19331/j.cnki.jxufe.2023.03.002.

[6] 汪三贵, 曾小溪. 后2020贫困问题初探[J]. 河海大学学报(哲学社会科学版), 2018, **20**(2): 7-13. DOI:10.3876/ j.issn.1671-4970.2018.02.002.
WANG S G, ZENG X X. A Study of Poverty after 2020[J]. *J Hohai Univ Philos Soc Sci*, 2018, **20**(2): 7-13. DOI: 10.3876/ j.issn.1671-4970.2018.02.002.

[7] 王小林, 冯贺霞. 2020年后中国多维相对贫困标准: 国际经验与政策取向[J]. 中国农村经济, 2020(3): 2-21. DOI: 11.1262.F.20200324.1717.002.
WANG X L, FENG H X. China's Multidimensional Relative Poverty Standards in the Post-2020 Era: International Experience and Policy Orientation[J]. *Chin Rural Econ*, 2020(3): 2-21. DOI:11.1262.F.20200324.1717.002.

[8] 檀学文. 走向共同富裕的解决相对贫困思路研究[J]. 中国农村经济, 2020(6): 21-36. DOI: 11.1262.

F.20200622.1108.014.
TAN X W. A Study on the Approach of Reducing Relative Poverty and Achieving Common Prosperity[J]. *Chin Rural Econ*, 2020(6): 21−36. DOI: 11.1262.F.20200622.1108.014.

[9] 李寻欢, 周扬, 陈玉福. 区域多维贫困测量的理论与方法[J]. 地理学报, 2020, **75**(4): 753−768. DOI: 10.11821/dlxb202004007.
LI X H, ZHOU Y, CHEN Y F. Theory and Measurement of Regional Multidimensional Poverty[J]. *Acta Geogr Sin*, 2020, **75**(4): 753−768. DOI: 10.11821/dlxb202004007.

[10] 董金鹏. 多维视角下我国相对贫困的测度与分析[D]. 南昌: 江西财经大学, 2021. DOI:10.27175/d.cnki.gjxcu.2021.000133.
DONG J P. Measurement and Analysis of Relative Poverty in China from Multidimensional Perspective[D]. Nanchang: Jiangxi University of Finance and Economics, 2021. DOI: 10.27175/d.cnki.gjxcu.2021.000133.

[11] 马鑫. 乡村振兴视域下相对贫困群体识别的随机森林模型研究[D]. 重庆: 重庆理工大学,2022. DOI: 10.27753/d.cnki.gcqgx.2022.000187.
MA X. Study on Random Forest Model for Identifying Relatively Poor Groups from the Perspective of Rural Revitalization[D]. Chongqing: Chongqing University of Technology, 2022. DOI: 10.27753/d.cnki.gcqgx.2022.000187.

[12] 曹睿, 廖彬, 李敏, 等. 基于XGBoost的在线短租市场价格预测及特征分析模型[J]. 数据分析与知识发现, 2021, **5**(6): 51−65. DOI: 10.11925/infotech.2096-3467.2020.1186.
CAO R, LIAO B, LI M, *et al*. Predicting Prices and Analyzing Features of Online Short-term Rentals Based on XGBoost[J]. *Data Anal Knowl Discov*, 2021, **5**(6): 51−65. DOI: 10.11925/infotech.2096-3467.2020.1186.

[13] 赵磊, 邓彤, 吴卓平. 基于数据挖掘的MOOC学习者学业成绩预测与群体特征分析[J]. 重庆高教研究, 2021, **9**(6): 95−105. DOI: 10.15998/j.cnki.issn1673-8012.2021.06.009.
ZHAO L, DENG T, WU Z P. The Prediction of Academic Achievement and Analysis of Group Characteristics for MOOC Learners Based on Data Mining[J]. *Chongqing High Educ Res*, 2021, **9**(6): 95−105. DOI: 10.15998/j.cnki.issn1673-8012.2021.06.009.

[14] LI Y J, FENG Y, QIAN Q. FDPBoost: Federated Differential Privacy Gradient Boosting Decision Trees[J]. *J Inf Secur Appl*, 2023, **74**: 103468. DOI: 10.1016/j.jisa.2023.103468.

[15] CHEN C X, GENG L W, ZHOU S. Retraction Note: Design and Implementation of Bank CRM System Based on Decision Tree Algorithm[J]. *Neural Comput Appl*, 2023, **35**(6): 4803. DOI: 10.1007/s00521-022-08194-1.

[16] MENG Q, KE G L, WANG T F, *et al*. A Communication-efficient Parallel Algorithm for Decision Tree[EB/OL]. arXiv Preprint: 1611.01276, 2016.

[17] 钱宇华, 王川杭, 王婕婷. 消除随机一致性的互信息及决策树算法[J]. 山西大学学报(自然科学版), 2022, **45**(5): 1206−1215. DOI: 10.13451/j.sxu.ns.2021016.
QIAN Y H, WANG C H, WANG J T. Mutual Information and Decision Tree Algorithm with Eliminating Random Consistency[J]. *J Shanxi Univ Nat Sci Ed*, 2022, **45**(5): 1206−1215. DOI: 10.13451/j.sxu.ns.2021016.

[18] GAO W Z, LI Z, CHEN Q S, *et al*. Modelling and Prediction of GNSS Time Series Using GBDT, LSTM and SVM Machine Learning Approaches[J]. *J Geod*, 2022, **96**(10): 1−17. DOI: 10.1007/s00190-022-01662-5.

[19] LI G Y, HAN G. The Behavior Analysis and Achievement Prediction Research of College Students Based on XGBoost Gradient Lifting Decision Tree Algorithm[C]// Proceedings of the 2019 7th International Conference on Information and Education Technology. New York: ACM, 2019: 289−294. DOI: 10.1145/3323771.3323803.

[20] CHEN T Q, GUESTRIN C. XGBoost: a Scalable Tree Boosting System[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 785−794. DOI: 10.1145/2939672.2939785.

[21] 廖彬, 王志宁, 李敏, 等. 融合XGBoost与SHAP模型的足球运动员身价预测及特征分析方法[J]. 计算机科学, 2022, **49**(12)195−204. DOI: 10.11896/jsjkx.210600029.
LIAO B, WANG Z N, LI M, *et al*. Integrating XGBoost and SHAP Model for Football Player Value Prediction and Characteristic Analysis[J]. *Comput Sci*, 2022, **49**(12)195−204. DOI: 10.11896/jsjkx.210600029.

[22] AKBAR S, ALI F, HAYAT M, *et al*. Prediction of Antiviral Peptides Using Transform Evolutionary & SHAP Analysis Based Descriptors by Incorporation with Ensemble Learning Strategy[J]. *Chemom Intell Lab Syst*, 2022, **230**: 104682. DOI: 10.1016/j.chemolab.2022.104682.

[23] LUNDBERG S M, LEE S I. A Unified Approach to Interpreting Model Predictions[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 4768−4777. DOI: 10.5555/3295222.3295230.

[24] 支港, 钟学燕, 王欣, 等. 基于Transformer的序列生成多标签文本分类[J]. 山西大学学报(自然科学版), 2023, **46**(1)10−19. DOI: 10.13451/j.sxu.ns.2022092.
ZHI G, ZHONG X Y, WANG X, *et al*. Transformer-based Sequence Generation for Multi-label Text Classification[J]. *J Shanxi Univ Nat Sci Ed*, 2023, **46**(1)10−19. DOI: 10.13451/j.sxu.ns.2022092.