山西大学学报(自然科学版)47(6):1190—1200,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024119

# 一种面向能源贫困识别的轻量可解释梯度提升树

王政[1],裔扬[2*],史颖[3,4],赵兴旺[4],吴晨旭[4]
(1.太原师范学院 智能优化计算与区块链技术山西省重点实验室,山西 晋中 030619;
2.扬州大学 信息工程学院,江苏 扬州 225127;
3.太原师范学院 计算机科学与技术学院,山西 晋中 030619;
4.山西大学 计算机与信息技术学院,山西 太原 030006)

**摘 要:**为解决传统梯度提升树方法在能源贫困识别时容易出现的训练不足、过拟合、可解释性差等问题,本文提出了一种面向能源贫困识别任务的轻量可解释梯度提升树方法。该方法首先剔除原始数据中的缺失值、异常值等噪声样本,根据特征关联分析后的样本梯度进行排序,以实现梯度提升树内部节点的分割,随后采用特征绑定技术加速训练过程,以实现模型的轻量化;其次,引入模型解释方法进行影响因素分析,量化不同特征对能源贫困识别的影响程度,增强了模型的可解释性。在典型的能源贫困识别数据集上的实验结果表明,与其他方法[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)、XGBoost(eXtreme Gradient Boosting)、GradientBoosting]相比,本文提出的轻量可解释模型的AUC(Area Under Curve)值达到99.61%,提高了0.2%~17.8%,具有较为明显的优势。

**关键词:**LightGBM(Light Gradient Boosting Machine)模型;能源贫困预测;特征关联分析;模型解释方法
**中图分类号:**TP391 **文献标志码:**A **文章编号:**0253-2395(2024)06-1190-11

# A Lightweight Interpretable Gradient Boosting Tree for Energy Poverty Identification

WANG Zheng[1], YI Yang[2*], SHI Ying[3,4], ZHAO Xingwang[4], WU Chenxu[4]
(1. *Shanxi Key Laboratory of Intelligent Optimization Computing and Blockchain Technology, Taiyuan Normal University, Jinzhong* 030619*, China;*
2. *School of Information Engineering, Yangzhou University, Yangzhou* 225127*, China;*
3. *College of Computer Science and Technology, Taiyuan Normal University, Jinzhong* 030619*, China;*
4. *School of Computer and Information Technology, Shanxi University, Taiyuan* 030006*, China*)

**Abstract:** In order to solve the problems of insufficient training, overfitting, and poor interpretability of the traditional gradient boosting tree method in identifying energy poverty, this paper designs a lightweight and interpretable gradient boosting tree for energy poverty identification. First, the noise samples such as missing values and outliers in the original data are eliminated, and the sample gradients after feature correlation analysis are sorted to realize the segmentation of internal nodes of the gradient boosting tree and to achieve the lightweight of the model. Then, the feature binding technology is used to accelerate the training process. Second, the model interpretation method is introduced to analyze the influencing factors to quantify the impact of different features on energy

**收稿日期:**2024-06-25;**接受日期:**2024-09-04
**基金项目:**国家自然科学基金(92371116)
**作者简介:**王政(1987—),男,山西汾阳人,博士,讲师,研究方向为计算社会学。E-mail:wz@tynu.edu.cn
**通信作者:**裔扬(YI Yang),E-mail:yiyang@yzu.edu.cn
**引文格式:**王政,裔扬,史颖,等.一种面向能源贫困识别的轻量可解释梯度提升树[J].山西大学学报(自然科学版),2024,47(6):1190-1200. DOI:10.13451/j.sxu.ns.2024119

poverty identification, which enhances the interpretability of the model. Experimental results on a typical energy poverty identification dataset show that compared with other methods [LR (Logistic Regression), KNN (K-Nearest Neighbor), SVM (Support Vector Machine), RF (Random Forest), CART (Classification and Regression Tree), XGBoost (eXtreme Gradient Boosting), GradientBoosting], the lightweight interpretable model proposed in this paper achieves an AUC (Area Under Curve) value of 99.61%, showing an improvement of 0.2% to 17.8%, and thus shows a more obvious advantage.

**Key words:** LightGBM model; energy poverty prediction; feature correlation analysis; model interpretation method

## 0 引言

能源贫困是可持续发展目标实现的一大阻碍,准确识别是解决该问题的首要关键环节,受到学术界的广泛关注[1]。传统计量回归方法以系数大小表征变量的重要性程度[2],准确性较低导致重复测量,同时解释性容易受到变量单位和数值量级影响,制约了精准能源扶贫的实施。

机器学习方法具备自动学习、识别和推理能力[3],能够提升能源贫困识别的精确性。随机森林的识别方法具有处理高维数据的能力被广泛应用于贫困识别[4],Arribas-Bel 等[5]提出了一种基于遥感技术和随机森林的测量方法,其可以快速分析大量数据,但容易受到噪声、环境等的影响,导致结果准确率不高。邻近算法(K-Nearest Neighbor,KNN)[6]在处理现实社会中的贫困问题上更简单直观,但其对异常值较为敏感。Reddy 等[7]通过提取每个贫困类别的特征子集,将其应用于集成分类器进行评估,并采用 LIME(Local Interpretable Model-Agnostic Explanations)进行解释,但 LIME 是一种局部可解释性方法,其缺乏全局解释能力。将支持向量机(Support Vector Machine,SVM)的分类模型应用于贫困家庭和非贫困家庭的数据分类领域,结果精确度为 88.64%[8],但在训练过程中对参数和核函数的选择敏感。Shen 等[9]采用决策树算法对贫困问题进行预测,平均准确率可达到 89%,但其处理连续特征困难。学界普遍认为基于梯度法的机器学习方法更适用于特征的识别[10]与预测[11]问题。Liu 等[12]建立梯度提升决策树(Gradient Boosting Decision Trees,GBDT)电力负荷预测模型,解决了预测的有效性问题,但 GBDT 通常输出单个变量,未能充分考虑输出变量之间的相关性,学习到的树结构中会出现冗余现象。Huang 等[13]基于 XGBoost(eXtreme Gradient Boosting)算法,运用极致梯度提升相对贫困人群的准确识别,准确率达到 81.9%,但存在过拟合的风险。

针对上述问题,本文提出了一种轻量可解释的梯度提升树,并将其应用于能源贫困识别。该方法首先剔除原始数据中的缺失值、异常值等噪声样本,根据特征关联分析后的样本梯度进行排序,以实现梯度提升树内部节点的分割,随后采用特征绑定技术加速训练过程;其次,引入模型解释方法进行影响因素分析,量化不同特征对能源贫困识别的影响程度,增强了模型的可解释性。本文工作主要集中在以下 3 个方面:

(1)利用梯度提升树建立了能源贫困识别模型,通过输入个人与家庭相关数据,判断能源贫困值与能源贫困线的差距,精确识别能源贫困。

(2)借助网格搜索方法对梯度提升模型的超参数加以调整,显著提升了模型的性能。实验结果表明,本研究提出的模型较多数传统模型具有显著的性能优势。

(3)引入模型解释方法进行影响因素分析,明晰不同数据特征和能源贫困存在直接的正向或负向相关,特征之间是否存在交互相关、特征对个体的影响等问题,为能源贫困的精准识别提供了决策参考。

## 1 轻量可解释模型的能源贫困识别方法

本文以中国家庭追踪调查数据(Chinese Family Panel Studies,CFPS)2018 年的能源贫困数据集为研究对象,提出一种融合轻量梯度提升模型与模型解释方法的能源贫困识别和特征分析方法,流程如图 1 所示。

核心步骤包括:(1)数据预处理,对数据集中的能源数据进行提取,处理原始数据中的缺


**图1** 能源贫困识别及特征分析模型示意图

**Fig. 1** Schematic diagram of energy poverty identification and feature analysis model

失值、异常值,运用方差分析法进行特征选择;(2)构建能源贫困识别模型,利用轻量梯度提升模型构建能源贫困识别模型,进行超参数调优,获得最优的模型架构;(3)融合模型解释方法,采用模型解释方法分析不同特征对能源贫困识别的影响程度。

### 1.1 特征选择

合理的特征组合能显著提高模型构建的效率和准确性。本文通过方差$p$值进行特征关联分析,在原始数据的基础上,排除无关特征的数据,选取有代表性、可靠性强和关联度高的指标。设样本数为$n$,因素有$k$个水平,每个水平的均值分别用$\mu_1,\mu_2,\cdots,\mu_k$表示具体步骤如下:

(1)提出假设。$H_0$:$\mu_1=\mu_2=\cdots=\mu_k$,$H_1$:$\mu_1,\mu_2,\cdots,\mu_k$特征值具有差异。

(2)计算统计量$F$。$F=\dfrac{\bar{S}_A}{\bar{S}_E}$,其中,$\bar{S}_A$为组间均方,$\bar{S}_E$为组内均方。

$$\bar{S}_A=\frac{S_A}{k-1}, \tag{1}$$

$$\bar{S}_E=\frac{S_E}{n-k}, \tag{2}$$

$S_T$为总误差平方和、$S_A$为组间平方和、$S_E$为组内平方和。

$$S_T=\sum_{i=1}^{k}\sum_{j=1}^{n_{ij}}(x_{ij}-\bar{\bar{x}})^2, \tag{3}$$

$$S_A=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i-\bar{\bar{x}})^2, \tag{4}$$

$$S_E=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2, \tag{5}$$

其中$\bar{\bar{x}}$代表所有特征值的均值,$\bar{x}_i$代表第$i$个观察值的均值,$n_i$代表$i$样本的数量,$x_{ij}$是$i$样本的$j$观察值。

(3)特征选择策略

当$F>F_\alpha$($F_\alpha$为特征均值)则说明均值间具有显著的差异,该因素对能源贫困有影响。反之,当$F<F_\alpha$,说明此因素对能源贫困无影响。

在假设零条件成立的前提下,$p$值表示观测到当前$F$值或更极端结果的概率。当$p<0.05$,则可以拒绝零假设,认为该特征与目标变量之间存在显著差异,即存在关联,则对应的特征被选择。反之,认为两者之间没有显著关联,丢弃相关的特征。

### 1.2 模型构建

用于能源贫困识别的数据集$S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,$n$表示第$n$个样本,$m$表示第$m$个特征,其中$x=(x_1,x_2,\cdots,x_m)$表示任意样本,包含家庭成员基本信息、家庭收入及家庭各类支出等特征,$y=(y_1,y_2,\cdots,y_n)$表示居民人均能源消费。

相较于简单的分类模型,集成模型往往可以取得更高的预测准确性。在高维输入特征和大规模数据下,轻量梯度提升模型相较于传统梯度提升模型,具有更高效的训练速度、更低的内存占用、更好的准确性和更好的特征并行化[14],因此本文构建轻量梯度提升模型进行能源贫困识别。

梯度提升模型$f(x)$的预测值是一组决策树模型$h_t(x)$的输出的总和:

$$f(x)=\sum_{t=1}^{T}h_t(x), \tag{6}$$

其中$T$代表决策树的数量。构建梯度提升模型的目的是寻找一个近似函数$\hat{f}$,该函数能够最小化损失函数$L[y,f(x)]$,如下所示:

$$\hat{f}=\arg\min_f E_{y,s}L[y,f(x)]。 \tag{7}$$

本文利用GOSS(Gradient-based One-Side Sampling)方法分割树的内部节点,这与传统梯度提升模型不同,后者通常依赖于信息增益。具体地,本文首先将样本按照梯度绝对值排序,子集$A$由梯度值较大的样本$a\%$构成,子集$B$由在剩余梯度较小的样本中随机抽取$b\times(1-a)\times 100\%$而形成,然后基于$A\cup B$的方差

增益 $V_j(d)$对这些样本进行分割。

$$V_j(d)=\frac{1}{n}\left[\frac{\left(\sum_{x_i\in A_1}g_i+\frac{1-a}{b}\sum_{x_i\in B_1}g_i\right)^2}{n_1^j(d)}+\frac{\left(\sum_{x_i\in A_r}g_i+\frac{1-a}{b}\sum_{x_i\in B_r}g_i\right)^2}{n_r^j(d)}\right], \tag{8}$$

其中 $A_1=\{x_i\in A:x_{ij}\leqslant d\}$, $A_r=\{x_i\in A:x_{ij}>d\}$, $B_1=\{x_i\in B:x_{ij}\leqslant d\}$, $B_r=\{x_i\in B:x_{ij}>d\}$, $g_i$表示模型在每次迭代中计算并输出损失函数的负梯度。

此外,本文还采用了独有的特征绑定技术,该方法能在保持训练精度的前提下加速训练过程。许多应用中存在高维且稀疏的输入特征,这些特征通常是相互排斥的(即这些特征不能同时为非零),本文利用特征绑定技术将这些特性捆绑到单一的特性包中。这些特征束和单个特征束的特征直方图都可以通过特征扫描算法构建。本文方法采用GOSS方法,基于方差增益分割内部节点,并使用独有的特征绑定技术进行输入特征降维,以达到缓解冗余贫困因素对模型构建的伤害。此外,作为一种基于决策树的模型,本文方法还具有抗多重共线性的鲁棒性的额外优势。

### 1.3 模型可解释性

能源贫困识别任务的数据通常维度较高且含有不少影响较小的变量,这些均可导致模型的训练速度与精度的降低。尽管轻量梯度提升模型具有更高的预测准确性,但在解释不同特征对识别结果的影响方面略有不足。事实上,理解不同特征与贫困类别的联系,区分不同因素对贫困的影响程度对于后续针对性解决贫困问题和调整优化模型意义非凡。

基于上述研究,围绕轻量梯度提升模型可解释性弱的不足,本文基于SHAP(SHapley Additive exPlanations)方法提出了一种模型解释方法,以对模型中影响居民人均能源消费指标$y$的因素进行解释分析,增强模型的可解释性。SHAP[15]是一种基于博弈论的方法,用于解释机器学习模型的输出,它将所有特征变量视为对模型输出的“贡献者”,本文通过计算特征贡献值,将样本中每个特征对模型最终输出结果的贡献进行量化。

能源贫困的识别不仅涵盖家庭生活中的能源和电力使用情况,还涉及家庭成员的个体特征[2],尽管机器学习方法能够提供判定结果,但其识别过程对于决策者来说仍然是一个“黑箱”。SHAP方法通过特征贡献值来评估不同特征对分类结果的影响程度,能够直观展示影响能源贫困的主要特征及其排序,从而为解决能源贫困问题提供更加丰富的决策参考。

在实际应用时,通常将输出值归因到每个特征的贡献值上来衡量其影响。具体而言,对于给定样本$x=(x_1,x_2,\cdots,x_m)$,$f_i$为给定样本的特征$i$的特征贡献值,表示对应样本的第$i$个特征对预测值的贡献值,其表示如下:

$$y=f_{\text{base}}+f_1+\cdots f_i+\cdots+f_M, \tag{9}$$

其中$f_{\text{base}}$为整个模型的基线,其一般为所有样本目标变量的均值。模型的识别值$y$可以表示为基线和各特征的贡献值之和,其公式如下:

$$y=f_{\text{base}}+f_1+\cdots f_i+\cdots+f_m。 \tag{10}$$

当$f_i>0$,说明该特征对预测值起到正向作用,反之$f_i<0$则起反向作用,且值越大对应的特征对预测结果的影响越显著。传统的特征重要性方法仅能显示出模型中特征的重要性,并不能明确各特征对预测结果的具体影响。因而,本文将轻量梯度提升模型与模型解释方法结合建模,不仅可以量化每个特征对预测结果的影响,还能标示其作用的正负方向,有助于进一步提升模型预测的可解释性和为消除能源贫困制定更合理的方案。

## 2 实验结果及分析

### 2.1 数据预处理

实验的数据集采用由北京大学社会科学研究所调查管理的2018年中国家庭追踪调查数据。该数据库是一个具有全国代表性的样本,覆盖了25个省、市或自治区,占中国人口的95%,数据集涵盖了家庭的社会经济指标、个人经济和非经济福祉指标,从中选取识别能源贫困选取的个人特征与家庭特征,获得影响能源贫困的因素共包括26个,依次用x1—x26表示(见表1)。将数值型能源人均年消费记为y0,类别型能源人均年消费记为y1,本文参照刘自

**表1** 能源贫困各项特征值描述性统计

**Table 1** Descriptive statistics of various feature values for energy poverty

| 符号 | 含义 | count | mean | std | min | max |
|---|---|---|---|---|---|---|
| x1 | 支出对数 | 13 468 | 10.75 | 1.00 | 0.00 | 14.52 |
| x2 | 家庭规模 | 13 468 | 3.55 | 1.91 | 1.00 | 21.00 |
| x3 | 家庭年人均纯收入 | 13 468 | 22 598.02 | 31 197.42 | 0.00 | 1 012 500.00 |
| x4 | 家庭年人均支出 | 13 468 | 25 317.32 | 35 148.96 | 0.00 | 1 614 900.00 |
| x5 | 现代燃料 | 13 468 | 1.21 | 6.02 | 0.00 | 77.00 |
| x6 | 房屋产权 | 13 468 | 0.83 | 0.38 | 0.00 | 1.00 |
| x7 | 房屋面积 | 13 468 | 22.53 | 67.32 | 0.00 | 3 000.00 |
| x8 | 管道设施 | 13 468 | 0.76 | 0.42 | 0.00 | 1.00 |
| x9 | 食品支出份额 | 13 468 | 0.36 | 0.48 | 0.00 | 18.00 |
| x10 | 衣着支出份额 | 13 468 | 0.05 | 0.08 | 0.00 | 3.00 |
| x11 | 生活用品支出份额 | 13 468 | 0.10 | 0.34 | 0.00 | 18.15 |
| x12 | 交通通信支出份额 | 13 468 | 0.09 | 0.14 | 0.00 | 7.10 |
| x13 | 文教娱乐支出份额 | 13 467 | 0.09 | 0.18 | 0.00 | 7.00 |
| x14 | 医疗保健支出份额 | 13 468 | 0.11 | 0.27 | 0.00 | 18.00 |
| x15 | 住房支出份额 | 13 468 | 0.14 | 0.33 | 0.00 | 12.59 |
| x16 | 其他支出份额 | 13 468 | 0.02 | 0.05 | 0.00 | 2.02 |
| x17 | 电费份额 | 13 468 | 0.03 | 0.05 | 0.00 | 3.00 |
| x18 | 燃气费份额 | 13 468 | 0.03 | 0.07 | 0.00 | 4.26 |
| x19 | 取暖费份额 | 13 468 | 0.01 | 0.02 | 0.00 | 0.80 |
| x20 | 年龄 | 13 468 | 50.08 | 15.37 | 11.00 | 95.00 |
| x21 | 性别 | 13 468 | 0.53 | 0.50 | 0.00 | 1.00 |
| x22 | 城乡 | 13 242 | 0.52 | 0.50 | 0.00 | 1.00 |
| x23 | 工作状况 | 13 266 | 0.75 | 0.43 | 0.00 | 1.00 |
| x24 | 婚姻 | 13 174 | 0.82 | 0.38 | 0.00 | 1.00 |
| x25 | 健康状况 | 13 382 | 0.68 | 0.47 | 0.00 | 1.00 |
| x26 | 受教育水平 | 13 456 | 1.96 | 1.03 | 1.00 | 4.00 |
| y0 | 数值型能源人均年消费 | 13 227 | 949.82 | 1 064.01 | 0.00 | 30 000.00 |
| y1 | 类别型能源人均年消费 | 13 468 | 0.52 | 0.50 | 0.00 | 1.00 |

敏等的研究[16],以中国居民每年能源消费最低线637.09元为能源贫困线,低于相对能源贫困线的取1,其余取0。

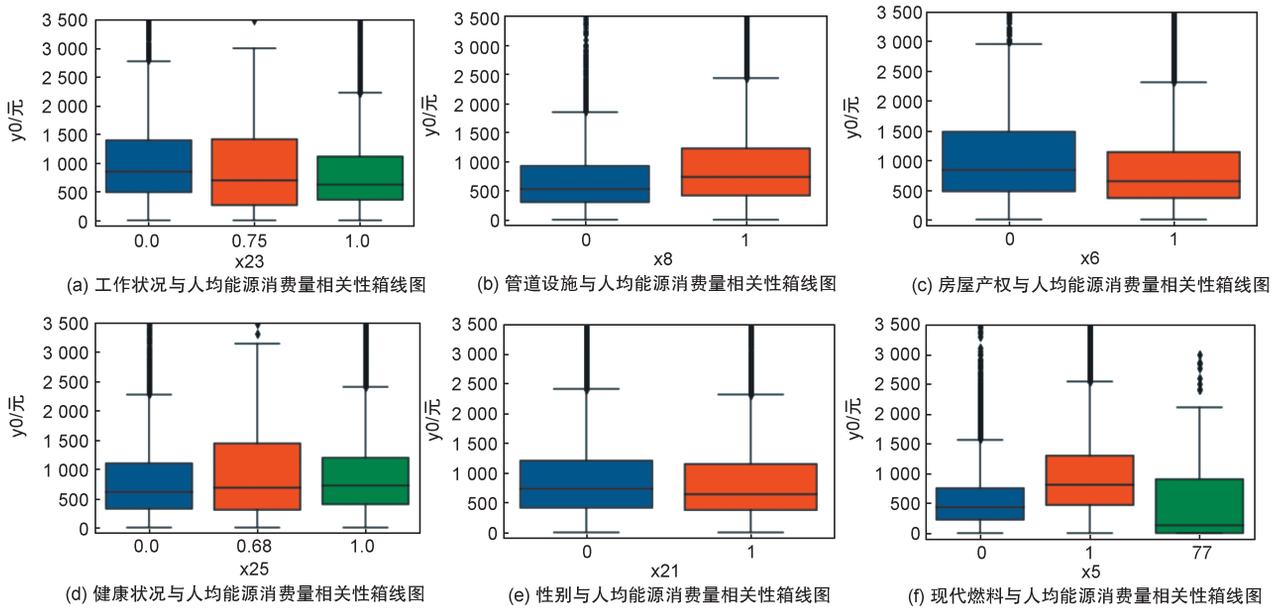
通过绘制特征值与人均能源消费量相关性箱线图对数据集特征进行探索及分析。图2中,箱体中央的线条表示数值的中位数,代表了样本的中间水平。箱体的宽度则反映了数据的变异程度,这种可视化方法能够帮助理解输入特征的分布和潜在异常值。

本文根据$p$值大小判断了检验特征与变量之间的相关性,并且根据计算得出的$p$值大小进行排序,$p$值越低显著水平越高,选取$p$值低于0.05的特征变量作为模型特征变量。根据$p$值在原始26个特征变量基础上,剔除x21、x15、x13、x5、x12、x19,最终选取的特征变量见表2。

## 2.2 模型参数优化

模型参数优化能够有效增强模型在训练与测试数据集上的性能及其对新数据的适应能力,同时可以有效预防过拟合与欠拟合现象的发生。本文采用了网格搜索技术,结合专家经验和穷举实验,为轻量梯度提升模型优化了四个核心参数。

为了确定每个参数的最优值,本文设定了一个包含四个参数各自候选值的参数网格,通过系统地遍历每一组参数组合,针对每一组组合分别训练模型,并通过一系列性能指标(如准确率、召回率等)来评估模型。通过比较不同参数组合下模型的性能,找到一组最优的参数配置,这组配置能够使得模型在训练集和测




(a) 工作状况与人均能源消费量相关性箱线图 (b) 管道设施与人均能源消费量相关性箱线图 (c) 房屋产权与人均能源消费量相关性箱线图

(d) 健康状况与人均能源消费量相关性箱线图 (e) 性别与人均能源消费量相关性箱线图 (f) 现代燃料与人均能源消费量相关性箱线图

**图2** 特征值与人均能源消费量的关系箱线图

**Fig. 2** Box plot of the correlation between feature values and per capita energy consumption

**表2** 特征变量的$p$值

**Table 2** $p$-values of features

| feature | f_classif | $p$ value |
|---|---|---|
| x2 | 6.130 5 | 0.000 0 |
| x19 | 5.718 4 | 0.000 0 |
| x18 | 2.699 7 | 0.000 0 |
| x4 | 2.245 8 | 0.000 0 |
| x1 | 2.099 3 | 0.000 0 |
| x11 | 2.077 1 | 0.000 0 |
| x7 | 2.065 9 | 0.000 0 |
| x26 | 1.859 2 | 0.000 0 |
| x22 | 1.814 5 | 0.000 0 |
| x24 | 1.625 2 | 0.000 0 |
| x10 | 1.574 9 | 0.000 0 |
| x17 | 1.514 6 | 0.000 0 |
| x3 | 1.451 3 | 0.000 0 |
| x16 | 1.447 7 | 0.000 0 |
| x14 | 1.429 1 | 0.000 0 |
| x23 | 1.310 4 | 0.000 0 |
| x20 | 1.232 5 | 0.000 0 |
| x8 | 1.160 7 | 0.000 0 |
| x6 | 1.131 8 | 0.000 3 |
| x25 | 1.074 5 | 0.023 4 |

试集上的表现达到最佳,从而有效提升模型的预测精度和泛化能力。这一过程确保了参数优化的全面性和系统性,有助于获得最优的模型配置。详见表3。

表4展示了不同参数在不同取值下的模型的预测效果。在学习率取0.3、迭代循环次数取200、决策树最大层数取3,随机抽样时选择的样本比例为1时,兼顾了模型运行性能与准确率,模型效果达到最优。

**表3** 模型最优参数及默认值

**Table 3** Optimal parameters and default values of model

| 参数名 | 默认值 | 参数最优值 | 参数含义 |
|---|---|---|---|
| learning_rate | 0.1 | 0.3 | 学习率 |
| n_estimators | 100 | 200 | 迭代循环次数 |
| max_depth | −1 | 3 | 决策树的最大层数 |
| Subsample | 1 | 1 | 随机抽样时选择的样本比例 |

### 2.3 与已有工作的实验对比

将本文方法LightGBM(Light Gradient Boosting Machine)与LR(逻辑回归,Logistic Regression)[17]、KNN[18]、SVM[19]、RF(Random Forest)[20]、CART(Classification and Regression Tree)[21]、XGBoost[22]、Gradient Boosting[23]等7种模型对比(详见表5),采用10折交叉验证方法进行性能评估。由表5可知,虽然本文提出的模型在fit_time和score_time花费的时间稍长一些,但其精确度和最优模型XGBoost仅差0.001 2,且相较于其他模型,准确率提高了1.13%~32.09%,recall提高了0.25%~32.08%,$F1$-score提高了0.06%~31.37%,以及ROC_AUC(Receiver Operating characteristic Curve_Area Under Curve)提高了0.02%~29.11%,在大多数指标上均好于其他模型,因此,本文提出的模

**表4** 模型参数取值及结果

**Table 4** Parameter values and results of model

| learning_rate/n_estimators/max_depth/Subsample | fit_time | score_time | accuracy | precision | recall | *F*1 | roc_auc |
|---|---|---|---|---|---|---|---|
| 0.1/100/—1/1(默认) | 0.105 9 | 0.008 8 | 0.972 3 | 0.970 4 | 0.976 3 | 0.973 3 | **0.996 2** |
| 0.3/200/3/1(参数最优值) | 0.110 4 | **0.007 3** | 0.973 0 | 0.970 6 | **0.977 4** | **0.974 0** | **0.996 2** |
| 0.2/200/3/1 | 0.081 0 | 0.009 1 | **0.973 1** | **0.974 7** | 0.973 3 | **0.974 0** | 0.995 1 |
| 0.1/200/3/1 | 0.076 5 | 0.009 2 | 0.969 0 | 0.971 7 | 0.968 3 | 0.970 0 | 0.994 5 |
| 0.3/100/3/1 | 0.046 8 | 0.008 2 | 0.970 7 | 0.971 4 | 0.972 0 | 0.971 7 | 0.994 9 |
| 0.2/100/1/1 | **0.028 2** | 0.007 7 | 0.910 2 | 0.894 0 | 0.937 5 | 0.915 2 | 0.971 0 |
| 0.3/200/2/1 | 0.055 4 | 0.008 4 | 0.972 1 | **0.974 7** | 0.971 2 | 0.973 0 | 0.994 5 |
| 0.3/200/1/1 | 0.078 4 | 0.009 0 | 0.969 0 | 0.971 7 | 0.968 3 | 0.970 0 | 0.994 5 |
| 0.3/200/3/0.5 | 0.077 6 | 0.009 3 | 0.972 5 | 0.974 5 | 0.972 2 | 0.973 3 | 0.995 1 |

**表5** 各类模型性能对比(%)

**Table 5** Comparison of performance across different models (%)

| 模型 | fit_time ↓ | score_time ↓ | accuracy ↑ | precision ↑ | recall ↑ | *F*1 ↑ | roc_auc ↑ |
|---|---|---|---|---|---|---|---|
| LightGBM | 0.110 439 | 0.007 28 | **0.972 974** | 0.970 624 | **0.977 389** | **0.973 983** | **0.996 162** |
| LR | 0.073 601 | 0.004 79 | 0.742 429 | 0.756 304 | 0.753 305 | 0.751 248 | 0.825 471 |
| KNN | **0.043 193** | 0.306 021 | 0.651 992 | 0.669 054 | 0.656 609 | 0.660 252 | 0.705 026 |
| SVM | 4.088 170 | 0.549 008 | 0.716 738 | 0.728 642 | 0.734 052 | 0.727 392 | 0.802 28 |
| RF | 1.524 530 | 0.047 778 | 0.931 989 | 0.932 306 | 0.940 805 | 0.935 327 | 0.986 293 |
| CART | 0.081 941 | **0.004 195** | 0.928 126 | 0.934 709 | 0.926 149 | 0.930 094 | 0.928 194 |
| GradientBoosting | 2.824 351 | 0.008 286 | 0.961 688 | 0.959 008 | 0.967 333 | 0.963 133 | 0.992 494 |
| XGBoost | 0.432 263 | 0.007 779 | 0.972 38 | **0.971 806** | 0.974 904 | 0.973 342 | 0.995 945 |

型在分类任务的全面性能上更为出色,能够更好地适应和应对复杂的数据分布和预测需求。

用准确率值绘制箱线图进行评价(如图3),研究结果显示,本文所提出的方法在识别精度上显著优于其他7种对比模型。

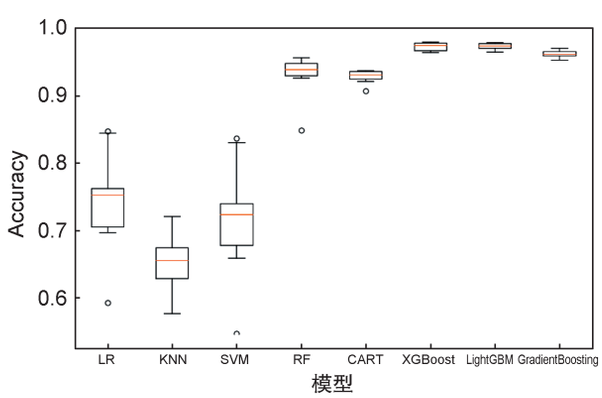



**图3** 不同算法准确率的箱型图比较

**Fig. 3** Comparison of box diagrams of accuracy of different algorithms

图4展示了8种模型识别能源贫困的ROC曲线,本文方法的ROC曲线最靠右上方,AUC值达到99.61%,相对于其他方法,提高了0.2%~17.8%,表明此模型的识别效果最佳。表5的实验结果显示,本文方法的预测准确率为97.58%,精确率为97.61%,召回率97.78%,*F*1值为97.69%,AUC值为99.79%,都在97%以上,与其他7种模型相比本文方法在精确率、召回率、*F*1值和AUC值均数最高表现更好,展示了模型的优越性。参数优化后本文方法在各项性能上与Gradient Boost比较接近。总之,本文构建的能源贫困识别模型在性能和准确性上表现优秀。

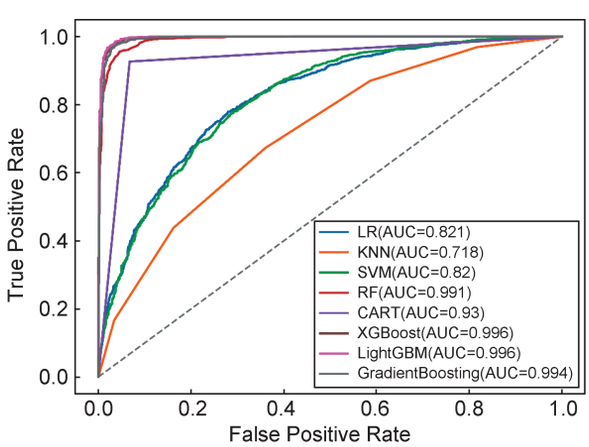
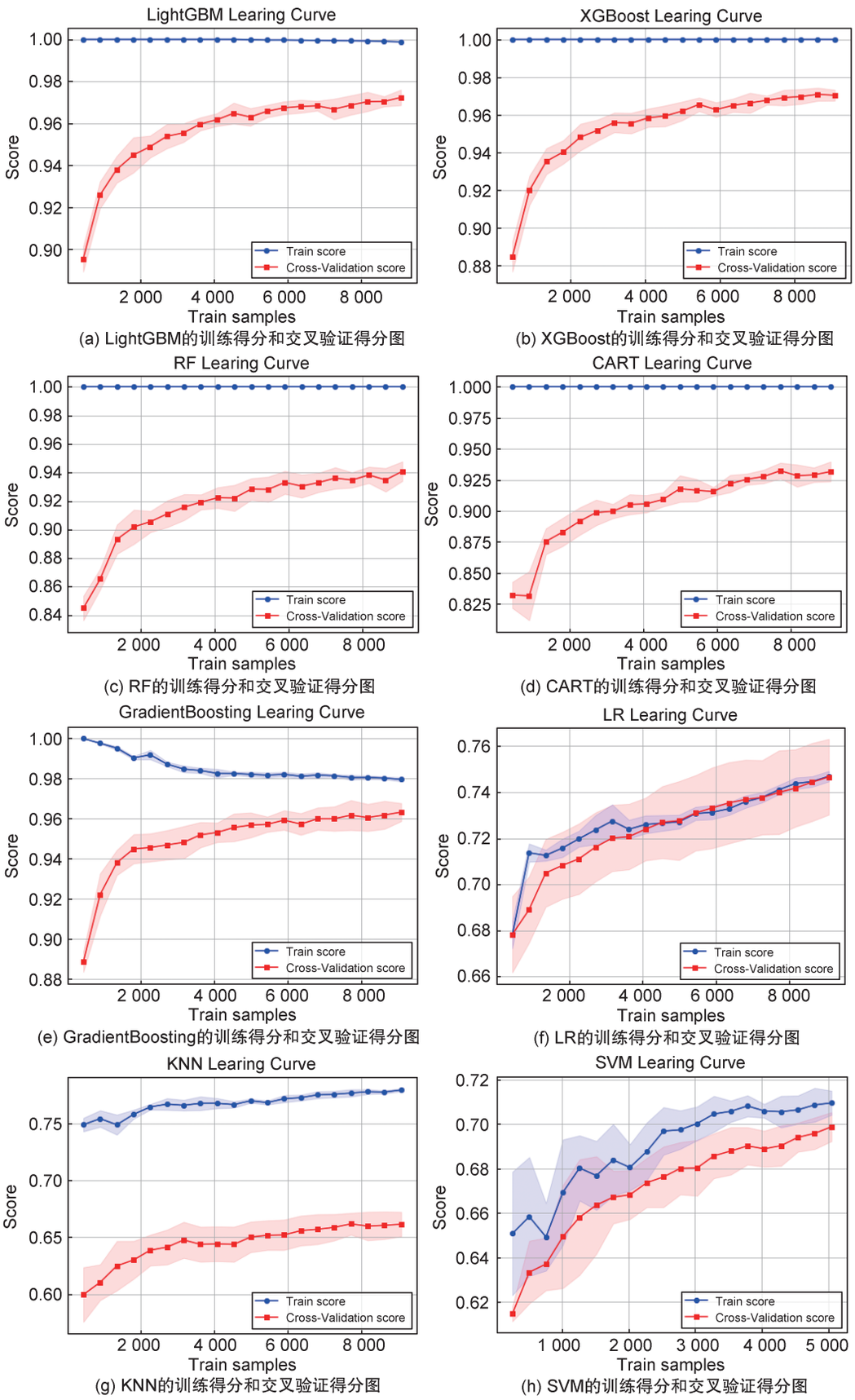



**图4** 各类模型ROC曲线对比图

**Fig. 4** Comparison chart of ROC curves for various models




(a) LightGBM的训练得分和交叉验证得分图 (b) XGBoost的训练得分和交叉验证得分图

(c) RF的训练得分和交叉验证得分图 (d) CART的训练得分和交叉验证得分图

(e) GradientBoosting的训练得分和交叉验证得分图 (f) LR的训练得分和交叉验证得分图

(g) KNN的训练得分和交叉验证得分图 (h) SVM的训练得分和交叉验证得分图

**图5** LightGBM与各类模型学习曲线对比图

**Fig. 5** Comparison of learning curves between lightGBM and various models

图5展示了8种算法在总体拟合趋势上不同的效果,无论是训练得分还是交叉验证得分本文方法都是最优的。本文方法与XGBoost、CART、RF这4种模型不受训练样本量影响,训练得分最好且稳定。而Gradient Boosting算法随着样本数量增大训练得分不断下降。

### 2.4 基于SHAP的模型解释分析

图6显示了SHAP值重要性排序图,将所有特征按SHAP值排序,显示了全局特征重要性分布。家庭支出全局重要度是最高的,其次分别为:燃气费份额、电费份额、取暖费份额、住房支出份额、家庭规模、支出对数,还有:房屋面积、家庭收入、食品支出份额。其中,排名第二的燃气费份额重要度只有排名第一家庭支出的一半左右,家庭收入后面特征的重要性很低。

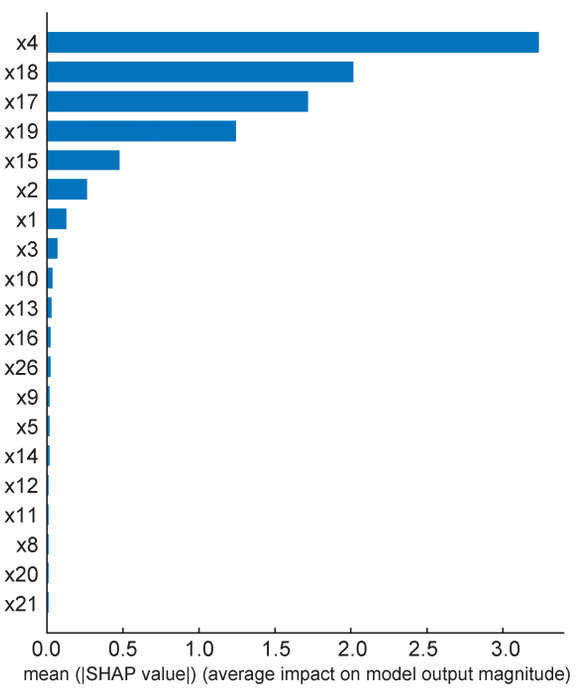



**图6** 特征重要性分析

**Fig. 6** Feature importance analysis

图7显示了总体特征图,不仅展示了个体的分布还能够反映出个体特征对能源贫困的影响程度。家庭人均支出越大对能源贫困有正向影响,而家庭人均支出越少则对能源贫困产生负向影响。与能源消费相关的特征是模型中影响最重要的特征,燃气费、电费和取暖费支出占比越大越有可能导致能源贫困。

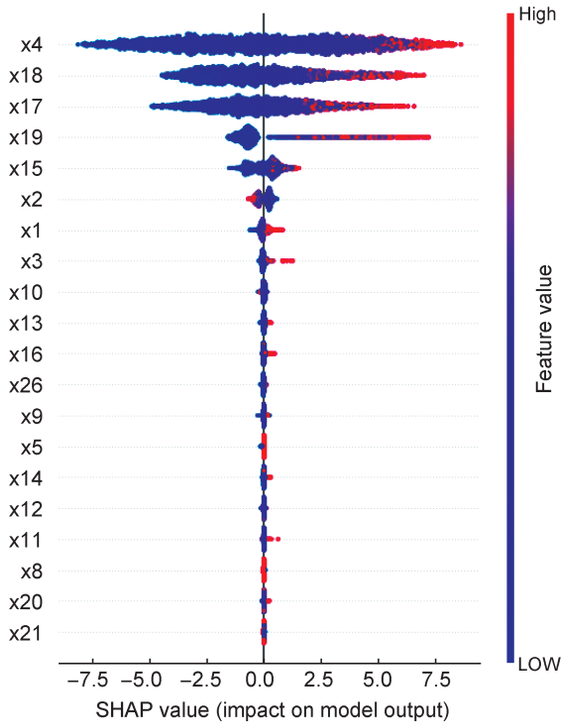



**图7** SHAP特征分类结果汇总图

**Fig. 7** Summary plot of SHAP feature classification results

## 3 结语

本文设计了一种轻量可解释的梯度提升模型,该方法根据特征关联分析实现了梯度提升树内部节点的分割,并采用特征绑定技术加速了模型的训练过程,随后,引入模型解释方法量化不同特征对能源贫困识别的影响程度,增强了模型的可解释性。与其他方法相比,本文提出的轻量可解释梯度提升模型在不同指标上均具有较为明显的优势,有效提高了能源贫困识别的准确性,并具备了能源贫困识别与解释的应用性。

**参考文献:**

[1] 宋轩,高云君,李勇,关庆锋,孟小峰.空间数据智能:概念、技术与挑战[J].计算机研究与发展,2022,**59**(2):255-263, DOI:10.7544/issn1000-1239.20220108.
SONG X,GAO Y J, LI Y, *et al*. Spatial Data Intelligence: Concepts, Technologies and Challenges[J]. *Comput Res Dev*, 2022, **59**(2): 255-263, DOI: 10.7544/issn1000-1239.20220108.

[2] 何可,朱信凯,李凡略.聚"碳"成"能":碳交易政策如何缓解农村能源贫困?[J].管理世界,2023,**39**(12):122-144. DOI: 10.19744/j.cnki.11-1235/f.2023.0143.

HE Ke, ZHU Xinkai, LI Fanliu. Gathering Carbon Into energy: How Can Carbon Trading Policies Alleviate Rural Energy Poverty?[J]. *Manage World*, 2023, **39**(12): 122−144, DOI: 10.19744/j.cnki.11-1235/f.2023.0143.

[3] 曹迪,谢敏杰,黄霆豪,等.改进蚁群算法的环境能源采集型WSN多目标路由研究[J].小型微型计算机系统, 2021, **42**(5): 1115−1120.
CAO D, XIE M J, HUANG T H, *et al*. Improved Ant Colony Algorithm for Multi-Objective Routing of Environmental Energy Harvesting WSN[J]. *J Chin Comput Syst*, 2021, **42**(5): 1115−1120.

[4] JAISWAL J K, SAMIKANNU R. Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression[C]//2017 World Congress on Computing and Communication Technologies (WCCCT). New York:IEEE, 2017: 65−68. DOI: 10. 1109/WCCCT. 2016. 25.

[5] ARRIBAS-BEL D, PATINO J E, DUQUE J C. Remote Sensing-based Measurement of Living Environment Deprivation: Improving Classical Approaches with Machine Learning[J]. *PLoS One*, 2017, **12**(5): e0176684. DOI: 10. 1371/journal. pone. 0176684.

[6] ZHANG S C, LI X L, ZONG M, *et al*. Efficient kNN Classification with Different Numbers of Nearest Neighbors [J]. *IEEE Trans Neural Netw Learn Syst*, 2018, **29**(5): 1774−1785. DOI: 10. 1109/TNNLS. 2017. 2673241.

[7] REDDY B, SRIKANYA K, VARSHINI M, *et al*. The Application of Machine Learning to the Task of Poverty Classification[C]//2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI). New York: IEEE, 2023: 1−4. DOI: 10. 1109/ACCAI58221. 2023. 10201169.

[8] YAN X C, LIN Z H, LIN Z Y, *et al*. A Novel Exploitative and Explorative GWO-SVM Algorithm for Smart Emotion Recognition[J]. *IEEE Internet Things J*, 2023, **10**(11): 9999−10011. DOI: 10. 1109/JIOT. 2023. 3235356.

[9] SHEN T P, ZHAN Z H, JIN L, *et al*. Research on Method of Identifying Poor Families Based on Machine Learning[C]//2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). New York: IEEE, 2021: 10−13. DOI: 10. 1109/IMCEC51613. 2021. 9482142.

[10] 王鑫, 廖彬, 李敏, 等. 融合LightGBM与SHAP的糖尿病预测及其特征分析方法[J]. 小型微型计算机系统, 2022, **43**(9): 1877−1885. DOI: 10.20009/j.cnki.21-1106/TP.2021-0114.
WANG X, LIAO B, LI M, *et al*. Combination of LightGBM and SHAP for Diabetes Prediction and Feature Analysis[J]. *J Chin Comput Syst*, 2022, **43**(9): 1877−1885. DOI: 10.20009/j. cnki. 21-1106/TP. 2021-0114.

[11] 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.
SHI Y, DING T Q, QI X B, *et al*. An Explainable Model for Identification and Early Warning of Relative Poverty [J]. *J Shanxi Univ Nat Sci Ed*, 2024, **47**(1): 155−165. DOI: 10.13451/j.sxu.ns.2023099.

[12] LIU S, CUI Y M, MA Y Z, *et al*. Short-term Load Forecasting Based on GBDT Combinatorial Optimization [C]//2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2). New York: IEEE, 2018: 1−5. DOI: 10. 1109/EI2. 2018. 8582108.

[13] HUANG W, LIU Y K, HU P Q, *et al*. What Influence Farmers' Relative Poverty in China: A Global Analysis Based on Statistical and Interpretable Machine Learning Methods[J]. *Heliyon*, 2023, **9**(9): e19525. DOI: 10. 1016/j. heliyon. 2023. e19525.

[14] Ke G, Meng Q, Finley T, *et al*. Lightgbm: A Highly Efficient Gradient Boosting Decision Tree[J]. *Adv Neural Inf Process Syst*, 2017, **30**, 3146−3154.

[15] WEN X, XIE Y C, WU L T, *et al*. Quantifying and Comparing the Effects of Key Risk Factors on Various Types of Roadway Segment Crashes with LightGBM and SHAP[J]. *Accid Anal Prev*, 2021, **159**: 106261. DOI: 10. 1016/j. aap. 2021. 106261.

[16] 刘自敏, 兰羽珩, 邓明艳, 等. 中国能源贫困的精准识别:基于等价尺度方法的分析[J]. 数量经济技术经济研究, 2023, **40**(2): 136−157.DOI:10.13653/j.cnki.jqte.2023.02.0004.
LIU Z M, LAN Y H, DENG M Y, *et al*. Accurate Identification of Energy Poverty in China: An Analysis Based on an Equivalent Scale[J]. *J Quant Tech Econ*, 2023, **40**(2): 136−157.DOI:10.13653/j.cnki.jqte.2023.02.0004.

[17] SIJM M P. Incremental Scannerless Generalized LR Parsing[C]//Proceedings Companion of the 2019 ACM SIGPLAN International Conference on Systems, Programming, Languages, and Applications: Software for Humanity. New York: ACM, 2019: 54−56. DOI: 10. 1145/3359061. 3361085.

[18] FAUZIAH, TIRO M A, RULIANA. Comparison of K-nearest Neighbor (K-NN) and Support Vector Machine (SVM) Methods for Classification of Poverty Data in Papua[J]. *ARRUS J Math App Sci*, 2022, **2**(2): 83−91. DOI: 10. 35877/mathscience741.

[19] NAVIAMOS M P, NIGUIDULA J D. A Study on De-

termining Household Poverty Status: SVM Based Classification Model[C]//Proceedings of the 3rd International Conference on Software Engineering and Information Management. New York: ACM, 2020: 79−84. DOI: 10. 1145/3378936. 3378969.

[20] ISHWARAN H, LU M. Standard Errors and Confidence Intervals for Variable Importance in Random Forest Regression, Classification, and Survival[J]. *Stat Med*, 2019, **38**(4): 558−582. DOI: 10. 1002/sim. 7803.

[21] 徐文庭, 殷昱煜, 王菊仙, 等. 基于CART与SlopeOne的服务质量预测算法[J]. 计算机集成制造系统, 2017, **23**(5): 1080−1090. DOI: 10.13196/j. cims. 2017.05.019. XU W T, YIN Y Y, WANG J X, *et al*. QoS Prediction Based on CART and SlopeOne[J]. *Comput Integr Manuf Syst*, 2017, **23**(5): 1080−1090. DOI: 10.13196/j. cims.2017.05.019.

[22] CHEN M H, LIU Q Y, CHEN S H, *et al*. XGBoost-based Algorithm Interpretation and Application on Post-fault Transient Stability Status Prediction of Power System[J]. *IEEE Access*, 2019, **7**: 13149−13158. DOI: 10. 1109/ACCESS. 2019. 2893448.

[23] HOU W L, SHI Q, LIU Y W, *et al*. State of Charge Estimation for Lithium-ion Batteries at Various Temperatures by Extreme Gradient Boosting and Adaptive Cubature Kalman Filter[J]. *IEEE Trans Instrum Meas*, 2024, **73**: 2504611. DOI: 10. 1109/TIM. 2023. 3346509.