第 4 卷◆第 4 期◆版本 1.0◆2021 年
文章类型：论文|刊号（ISSN）：2630-4759 /（中图刊号）：270GL018

基于机器学习模型的宁夏六盘山地区农户多维贫困影响因素研究

孟莹
北方民族大学 数学与信息科学学院
DOI:10.12238/ej.v4i4.737

[摘　要] 本文利用宁夏六盘山地区农户家庭调研数据,运用A-F双界法对多维贫困家庭进行识别,应用Logistic回归、决策树及随机森林等机器学习模型,从农户户主特征及家庭资本两方面探讨多维贫困的影响机制。结果表明：户主的受教育年限、户主性别、家庭劳动力数量及家庭外出务工人数对农村家庭的多维贫困具有显著的负向影响,户主年龄、家庭人口数、房屋用料基础对农村家庭多维贫困具有显著的正向效应。基于以上结论,提出完善农村贫困测度体系、关注农户生活水平、提高农户受教育程度、加大对农户就业扶持力度及出台相关人才引进政策等缓解多维贫困的政策建议。
[关键词] 多维贫困；Logistic回归；决策树；随机森林；影响因素
中图分类号：F323　文献标识码：A

Research on Influencing Factors of Multi-dimensional Poverty of Farmers in Liupan Shan Area of Ningxia Based on Machine Learning Model

Ying Meng
School of Mathematics and Information Science, North Minzu University

[Abstract] Based on the survey data of farmers' families in Liupanshan area, Ningxia, this paper uses A-F double-bound method to identify multi-dimensional poor families, and applies machine learning models such as Logistic regression, decision tree and random forest to explore the influencing mechanism of multi-dimensional poverty from the characteristics of farmers' heads of households and family capital. The results show that the years of education of the head of household, the head of household's sex, the number of family labor force and the number of family migrant workers have significant negative effects on the multidimensional poverty of rural families, while the age of the head of household, the number of family population and the basis of housing materials have significant positive effects on the multidimensional poverty of rural families. Based on the above conclusions, this paper puts forward some policy suggestions to alleviate multidimensional poverty, such as perfecting rural poverty measurement system, paying attention to farmers' living standards, improving farmers' education level, increasing employment support for farmers, and introducing relevant talent introduction policies.
[Key words] multidimensional poverty; Logistic regression; Decision tree; Random forest; influencing factor

引言

我国绝对贫困基本消除后,扶贫工作将由实现“两不愁、三保障”目标向应对缓解发展不平衡、不充分的多维相对贫困转变。了解导致农户多维贫困的因素,对症下药,不仅对缓解农民多维贫困起着重要作用,也是解决该地区多维贫困问题最直接最有效的减贫路径。

在对多维贫困的影响因素研究中,王艳慧[1]等运用最小二乘误差模型和空间计量分析模型识别了我国村庄的贫困类型和贫困差异,研究发现个体贫困的主要影响因素包括道路建设条件不发达、自然灾害频发、收入水平低、劳动条件差。侯亚景[2]通过多层回归模型实证分析了中国农村家庭长期多维贫困及不平等的影响因素。黄善林[3]通过PSM倾向得分匹配法分析了东北地区农户农地经营与农地流转对多维贫困的影响机制。本文以乡村振兴战略为背景设定多维贫困测度指标,从农户户主特征及家庭资本即农户自身角度出发来探讨多维贫困的影响因素,基于宁夏六盘山片区实地调研获取数据,运用机器学习模型进行分析,以取得更好的实证效果。

1 数据来源

本文数据来源于2020年9-10月课题组对泾源、西吉两县的农户调查。按照简单随机抽样的方法,依据“乡镇—村组—村民小组”的抽样流程,共发放问卷355份,其中泾源县205份,西吉县150份。回收有效问卷321份,有效回收率91.7%。本次调查采取面对面访谈并由调查人员

表 1 变量描述性统计

变量类型	变量名称	变量赋值	最小值	最大值	均值	标准差
因变量	贫困状态	贫困=1,非贫困=0	0	1	0.798	0.402
户主特征	性别	男=1,女=0	0	1	0.427	0.495
	年龄	实际年龄	15	84	47.707	10.579
	受教育年限	实际受教育年限	0	19	5.897	5.153
	政治面貌	党员=1,其他=0	0	1	0.069	0.253
	民族	汉族=1,少数民族=0	0	1	0.371	0.484
	就业情况	务农、务工=0,无劳动能力、待业=1	0	1	0.168	0.375
家庭资本	耕地面积	人均耕地面积	0	100	5.504	11.221
	劳动力数量	家庭实际劳动力人数	0	6	1.988	1.101
	房屋用料基础	钢混=1,砖混=2,砖木=3,土坯=4,石窑=5	1	5	2.050	0.809
	贫困类别	建档立卡贫困户、低保户、五保户=1,以上都不是=0	0	1	0.611	0.488
	家庭规模	家庭实际人口数	1	10	4.953	1.620
	家庭外出务工人数	家庭实际务工人数	0	4	0.897	0.894
	拥有资产情况	家庭中含有交通工具、农用机械设备总和小于 1000 元,赋值为 1	0	1	0.620	0.486

表 2 各模型预测结果

		多维贫困	非多维贫困	总计
Logistic 回归	多维贫困	77	4	81
	非多维贫困	11	5	16
	总计	88	9	97
决策树回归	多维贫困	72	9	81
	非多维贫困	8	8	16
	总计	80	17	97
随机森林回归	多维贫困	77	4	81
	非多维贫困	4	12	16
	总计	81	16	97

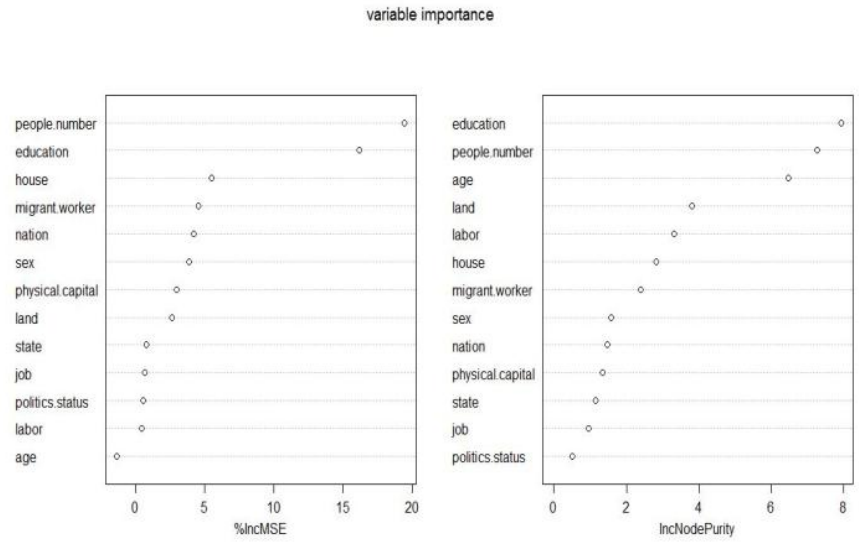

图 1 影响因素重要性排名度量

填写问卷的形式进行，问卷主要针对户主或熟悉家庭情况的其他家庭成员，其内容涉及农户家庭基本信息、日常状况、所享受的扶贫政策等。

2 研究方法

2.1多维贫困家庭的测算

本文选用A-F双界限法，从教育、健康及生活水平维度测度宁夏六盘山地区农户多维贫困现状，识别出多维贫困家庭，具体分两个步骤：第一步，测算各省份农村地区在某一贫困指标上的贫困程度是否超过某一剥夺临界值（即贫困线）；第二步，根据各贫困指标的权重和相对贫困数值，测算多维贫困指数。

2.2农户户主特征、家庭资本对多维贫困影响研究方法

(1)Logistic回归模型。通过Logit变换后，利用一般线性回归模型建立解释变量与被解释变量之间的多元分析模型。(2)决策树模型。决策树模型思想是自上而下在树的内部结点采用“递归”方法对属性值进行比较，即选择一个分类能力最好的变量把样本集分为多个子集，形成了分支，接着再对每个分支进行递归处理，选择下一个分类能力最好的变量进行分类，直到在该节点无法分类则停止递归，最终在叶子节点处得到最终分类结果。(3)随机森林模型。随机森林(Random Forest，以下简称RF)模型以决策树为基础，利用集成学习思想，“随机”构建多个独立且同分布的决策树并组成“森林”，遵从“少数服从多数”原则进行分类决策。

2.3变量设置

根据多维贫困测算结果，本文贫困状态为因变量，从户主特征和家庭资本两方面作为控制变量进行实证分析。其各变量描述性统计如表1所示。

3 实证分析

3.1各模型运行结果对比

由表2可知，各模型预测效果良好，且对是多维贫困农户预测效果均优于非多维贫困农户的预测效果。

3.2结果分析

3.2.1影响因素对比分析。下图1为RF依据不同方法给出的各变量重要性度

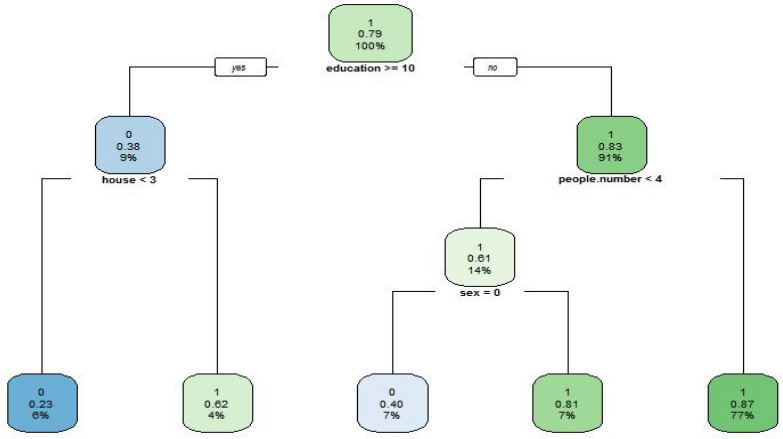

图 2 分类回归树

表 3 模型分类性能

评价标准 / 分类模型	Accuracy (%)	Precision (%)	Recall (%)	AUC	F1—Score
Logistic 回归	0.845	0.875	0.951	0.625	0.911
随机森林	0.918	0.951	0.951	0.998	0.951
决策树	0.825	0.900	0.889	0.631	0.894

量(降序排列)。可以看出,依据IncMSE排序前10位影响因素最大为人口数量,就业情况影响因素最小。排序前10个变量与 Logistic回归模型显著变量相比,有6个相同。对比两种度量方式,可以看到,家庭人口数量、户主受教育程度和房屋用料基础、家庭外出务工人数在两种方法中均排名靠前,说明以上4种因素对农户家庭多维贫困具有重要影响。较为遗憾的是,RF模型无法给出各变量具体的影响程度。

图2是分类回归树(CART),可以看出,区分度最好的特征依次是户主受教育程度、房屋用料基础、家庭人口数量、户主性别,与Logistic回归相比,Logistic回归的显著性因素包含了决策树算法结果的所有重要因素。

综上,对比3个模型影响因素结果发现,关于重要因素的模型运算结果相差不大,家庭人口数量、户主受教育程度、房屋用料基础、户主性别、家庭外出务工人数等因素是影响家庭多维贫困的主要因素。

3.2.2模型分类预测性能对比。在计算评价指标结果时,把是多维贫困农户作为正类别,把非多维贫困农户作为负类别,利用Logistic 回归、RF及决策树对预测样本集进行分类预测。

表3是各模型分类预测性能指标对比表格。总体来说,F1-Score作为精确率与召回率的综合指标,均大于或等于0.9,表明模型分类效果均较好;同时,AUC值越高说明模型的分类效果越好。对比评价指标平均值发现,从正确率(Accuracy)看,随机森林模型最高;从查准率(Precision)看,随机森林最高;从查全率(Recall)看,随机森林和Logistic回归一致;从F1-Score看,随机森林最高。对比Logistic回归、随机森林和决策树构建的多维贫困影响多因素分析模型,从模型准确率、精确率、AUC、F1-Score等指标看,随机森林表现最优,能够最大限度的对多维贫困农户进行正确分类,在对宁夏六盘山地区农户多维贫困样本判别效果最优。

4 结论与启示

基于多维贫困理论,以乡村振兴战略为背景构建农户多维贫困指标体系,运用A-F双界法对宁夏六盘山地区农户进行多维贫困的测算。运用Logistic回归模型、决策树模型及随机森林模型,探讨了户主特征和家庭资本对农户多维贫困的影响,主要得出以下结论:(1)农户的户主特征影响。户主的受教育年限对农村家庭的多维贫困具有显著的负向影响,户主的受教育年限越多,家庭陷入多维贫困的可能性越小;户主年龄对农村家庭多维贫困具有显著的正向效应,户主年龄越大,农户家庭越容易陷入多维贫困状态。(2)农户的家庭资本影响。家庭人口数对农村家庭多维贫困具有显著的正向影响,家庭人口数越多,农村家庭多维贫困的发生率越高;房屋用料基础对农村家庭多维贫困具有显著的正向影响,房屋质量越低,多维贫困发生的可能性越大;家庭劳动力数量和家庭外出务工人数对于农村家庭的多维贫困具有显著的负向影响,家庭劳动力数量和家庭外出务工人数越多,家庭陷入多维贫困的可能性越小。对此,我们可以得出,减贫不仅要关注农户的生活水平,更要提高农户的受教育程度,不仅要注重收入的提高更要重视可行能力的提高,例如加大对农户的就业扶持力度以及人才引进力度,从而更好地帮助农户摆脱多维贫困。

[基金项目]

北方民族大学研究生创新资助项目(YCX20104)。

[参考文献]

[1]王艳慧,陈烨烽,迟瑶,等.中国贫困村致贫因素分析及贫困类型划分(英文)[J].Journal of GeographicalSciences,2018,28(10):1444-1466.

[2]侯亚景.中国农村长期多维贫困的测量、分解与影响因素分析[J].统计研究,2017,34(11):86-97.

[3]黄善林,孙怡平,余志刚.农地经营与农地流转对多维贫困的影响研究——基于东北地区典型贫困县的农户调查[J].农林经济管理学报,2020,19(6):735-744.

作者简介:

孟莹(1997--),女,汉族,辽宁阜新人,硕士研究生,研究方向:经济与社会统计。