

健康风险视角下我国潜在返贫风险人群影响因素探索及预测研究

——基于CHFS数据

彭 朋, 周艺林, 庞 艳, 汪清雅*

重庆市结核病防治所, 重庆

收稿日期: 2026年8月20日; 录用日期: 2026年9月12日; 发布日期: 2026年9月21日

摘 要

目的: 识别潜在返贫风险高危因素, 精准识别潜在返贫风险家庭可以为潜在返贫风险人群的早期识别、超前干预提供理论依据。方法: 基于中国家庭金融调查数据库2019年数据, 利用相关性分析以及LASSO回归结合多因素Logistic回归识别潜在返贫危险因素, 利用十折交叉验证训练Logistic回归和XGBoost模型, 比较两模型对于潜在返贫风险家庭的预测效能。结果: 相关性分析与LASSO回归筛选变量后纳入17个潜在风险变量, 多因素Logistic回归显示潜在返贫风险与健康状况、资产、收入、教育、户籍类型、大病保险、政府帮扶、脱贫户等因素有关。Logistic模型的F1值为0.609、AUC值为0.734, XGBoost模型的F1值为0.943、AUC值为0.948, XGBoost模型预测效能优于Logistic回归, 收入、脱贫户、政府帮扶、户籍、存款是对模型贡献度最高的五个变量。结论: XGBoost模型能较为准确地预测潜在返贫风险, 该模型能早期识别潜在返贫风险人群, 为相关部门实施早期干预提供理论依据, 助力全面推进乡村振兴。

关键词

潜在返贫人群, 风险预测, XGBoost算法, LASSO回归, 影响因素

Exploration and Prediction Study of Influencing Factors of the Population at Risk of Potential Return to Poverty in China under the Perspective of Health Risk

—Based on CHFS Data

*通讯作者。

文章引用: 彭朋, 周艺林, 庞艳, 汪清雅. 健康风险视角下我国潜在返贫风险人群影响因素探索及预测研究[J]. 统计学与应用, 2026, 15(9): 205-216. DOI: 10.12677/sa.2026.159219

Peng Peng, Yilin Zhou, Yan Pang, Qingya Wang*

Chongqing Municipal Institute of Tuberculosis, Chongqing

Received: August 20, 2026; accepted: September 12, 2026; published: September 21, 2026

Abstract

Objective: To identify high-risk factors associated with the potential risk of return to poverty; accurately identifying households at such risk provides a theoretical basis for the early detection of—and proactive intervention for—the affected population. **Methods:** Based on data from the 2019 China Household Finance Survey, this study employed correlation analysis, LASSO regression, and multivariate logistic regression to identify risk factors. Ten-fold cross-validation was used to train logistic regression and XGBoost models, and their predictive performance regarding households at risk of return to poverty was compared. **Results:** Seventeen potential risk variables were selected for inclusion following correlation analysis and LASSO regression. Multivariate logistic regression indicated that the risk of return to poverty was associated with factors such as health status, assets, income, education, household registration type, critical illness insurance, government assistance, and status as a household that had previously escaped poverty. The logistic model yielded an F1 score of 0.609 and an AUC of 0.734, whereas the XGBoost model achieved an F1 score of 0.943 and an AUC of 0.948; the XGBoost model demonstrated superior predictive performance. Income, status as a household that had escaped poverty, government assistance, household registration type, and savings were the five variables contributing most significantly to the model. **Conclusion:** The XGBoost model can accurately predict the risk of return to poverty. It enables the early identification of at-risk populations, provides a theoretical basis for relevant authorities to implement early interventions, and supports the comprehensive advancement of rural revitalization.

Keywords

Potential Return to Poverty, Risk Prediction, XGBoost Algorithm, LASSO Regression, Influencing Factors

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

中国提前十年完成联合国 2030 年可持续发展议程的减贫目标，为全球减贫进程的推进作出了突出贡献。然而，各种纷繁复杂的风险冲击不断来袭，返贫风险问题依然存在。据统计，2020 至 2021 年全世界约有 0.88 亿~1.5 亿人重返贫困[1]。人民日报报道，2025 全国保障城乡低保对象 3910.4 万人、特困人员 486.1 万人，惠及 1189.7 万困难残疾人和 1664 万重度残疾人[2]。乡村振兴阶段，在脱贫攻坚的成果上，实现精准扶贫与乡村振兴有效衔接的基础和前提是不发生规模性返贫[3]。由此可见，脱贫摘帽不是奋斗的终点，而是新生活、新气象的起点。后扶贫时代贫困的治理视野应更加宽阔，应逐步实现由绝对贫困到相对贫困，由解决短期问题到建立长效脱贫机制的重点转移。

疾病或是后扶贫时代返贫致贫的首要原因。国家乡村振兴局建档立卡数据显示，近年来我国建档立

卡贫困户中因病致贫、因病返贫的约占 40%，出现了“广覆盖”的基本医疗保险与“高比例”的因病致贫并存的现象[4]。同时，北京同心圆慈善基金会发现，即使在基本医保和大病保险的保障下，对极重病人进行托底救助依然存在一定缺口，其因病返贫致贫风险尤为突出[5]。而健康风险发生是因病返贫、因病致贫链条的逻辑起点，因病返贫致贫本质上是健康风险冲击下的一种支出型贫困状态[6][7]。因此，探索如何及早识别健康风险下的潜在返贫人群，并帮助其有效抵御健康风险发生带来的冲击，从而遏制因病返贫致贫风险势必成为我国民生工程建设的重点环节。

长久以来，贫困问题都是学术界探讨的热点，国内外对于贫困影响因素的研究已经较为全面。现有研究已从家庭内部状态、家庭成员情况、家庭所处环境因素等家庭微观方面进行探讨，对于家庭外部因素多从地域、家庭社会网络、政府社会支持等方面探究影响因素[8]-[12]。可见，对于部分脱贫户、低保户，或是边缘户而言，他们自身抵御风险能力不强、地区内生发展能力不足，更易受到外界冲击而返贫致贫。然而，对于潜在返贫风险人群的预测类则相对较少。现有研究主要从贫困脆弱性测度、返贫预测、返贫预警系统展开研究，研究方法多选择传统模型算法、机器学习或深度学习，亦有研究综合比较多种模型预测效能[13]-[17]。

但就现有研究而言，该领域依然存在深入研究必要性。一是大多研究贫困划分标准按较为单一，多为年均收入低于绝对贫困线[18]，忽略了家庭外部因素所带来的影响，难以准确反映现阶段我国潜在返贫风险人群。二是研究针对健康风险这一返贫重要原因进行深入分析和探讨，对因病返贫的风险预测尚不完善，三是研究所用数据大多为小范围家庭调查数据，或是地区监测数据，代表性还有待提升。

本研究以健康风险为主视角，选用中国家庭金融调查数据库(China Household Finance Survey, CHFS)，全面筛查潜在影响因素，利用相关性分析及 LASSO 回归降维剔除高相关性变量，以多因素 Logistic 回归探究我国潜在返贫风险人群的影响因素、建立 Logistic 预测模型与 XGBoost 模型综合比较选取最优风险模型，构建健康风险下的潜在返贫风险预测模型，运用 SHAP 使模型可视化。以期精准探究我国潜在返贫风险人群的高危因素，精准预测潜在返贫风险人群，为潜在返贫风险人群的早期识别、超前干预提供理论依据。

2. 研究方法

2.1. 数据来源

数据选取中国家庭金融调查(China Household Finance Survey, CHFS) 2019 年最新公开数据，CFPS 西南财经大学中国家庭金融调查与研究中心自 2010 年起，截至目前共完成 7 轮中国家庭金融调查，主要内容包括：住房资产与金融财富、负债与信贷约束、收入与消费、社会保障与保险、代际转移支付、人口特征与就业以及支付习惯等相关信息，对家庭经济、金融行为进行了全面细致的刻画。数据涵盖了全国 29 个省份，355 个县(区、县级市)，包含 40,011 户家庭，在全国具有良好的代表性。CFPS2019 年调查数据包含“家庭数据”与“个人数据”两个数据包(下载链接：<https://chfser.swufe.edu.cn/datas/Products/Datas/DataList>)，其中“家庭数据”共含有 34,643 例样本、2656 个变量，“个人数据”包含 107,008 例样本、423 个变量。

2.2. 数据处理

本研究从家庭经济保障、家庭支出、健康风险、社会保障四个维度全面筛选两个数据包，共筛选出 68 个可能变量，将缺失大于 30% 的变量予以剔除，变量缺失小于 30% 的删除缺失样本，最终将变量合并转化剔除相似变量得到 26,625 例样本 28 个变量。由于家庭经济类数据离散程度过高且存在极端值，故根据原始数据特征及参考相关文献将这类变量转换为有序多分类变量，具体变量赋值见表 1。

Table 1. Variable assignment description table
表 1. 变量赋值说明表

变量	指标	赋值说明	变量	指标	赋值说明
X1	家庭是否从事工商业活动	是 = 1, 否 = 0	X2	家庭住房属性	自有房 = 1, 非自有房 = 0
X3	家庭拥有住房数量	——	X4	家庭是否有拆迁经历	是 = 1, 否 = 0
X5	家庭是否拥有汽车	是 = 1, 否 = 0	X6	家庭是否有积蓄	是 = 1, 否 = 0
X7	户主是否使用智能机	是 = 1, 否 = 0	X8	户主户口类型	农村 = 1, 城镇 = 2
X9	家庭人数	——	X10	家庭成员不健康人数比	——
X11	家庭成员最高学历	未上学 = 1, 小学 = 2, 初中 = 3, 高中 = 4, 大学 = 5, 研究生及以上 = 6	X12	家庭幸福感	非常幸福 = 1, 幸福 = 2, 一般 = 3, 不幸福 = 4, 非常不幸福 = 5
X13	家庭是否借出资金	是 = 1, 否 = 0	X14	家庭是否贷款	是 = 1, 否 = 0
X15	家庭去年固定支出	0~1 万 = 1, 1 万~2 万 = 2, 2 万~3 万 = 3, 3 万~4 万 = 4, 4 万~5 万 = 5, 5 万~6 万 = 6, 6 万~7 万 = 7, 7 万~8 万 = 8, 8 万~9 万 = 9, 9 万~10 万 = 10, 10 万以上 = 11	X16	家庭去年非固定支出	0~0.5 万 = 1, 0.5 万~1 万 = 2, 1 万~1.5 万 = 3, 1.5 万~2 万 = 4, 2 万~2.5 万 = 5, 2.5 万~3 万 = 6, 3 万~3.5 万 = 7, 3.5 万~4 万 = 8, 4 万~4.5 万 = 9, 4.5 万~5 万 = 10, 5 万以上 = 11
X17	家庭去年工资性收入	0~1 万 = 1, 1 万~2 万 = 2, 2 万~3 万 = 3, 3 万~4 万 = 4, 4 万~5 万 = 5, 5 万~6 万 = 6, 6 万~7 万 = 7, 7 万~8 万 = 8, 8 万~9 万 = 9, 9 万~10 万 = 10, 10 万以上 = 11	X18	家庭去年人均年收入	0~0.5 万 = 1, 0.5 万~1 万 = 2, 1 万~1.5 万 = 3, 1.5 万~2 万 = 4, 2 万~2.5 万 = 5, 2.5 万~3 万 = 6, 3 万~3.5 万 = 7, 3.5 万~4 万 = 8, 4 万~4.5 万 = 9, 4.5 万~5 万 = 10, 5 万以上 = 11
X19	家庭耐用品价值	0~0.5 万 = 1, 0.5 万~1 万 = 2, 1 万~1.5 万 = 3, 1.5 万~2 万 = 4, 2 万以上 = 5	X20	家庭存款	0~0.5 万 = 1, 0.5 万~1 万 = 2, 1 万~1.5 万 = 3, 1.5 万~2 万 = 4, 2 万~2.5 万 = 5, 2.5 万~3 万 = 6, 3 万~3.5 万 = 7, 3.5 万~4 万 = 8, 4 万~4.5 万 = 9, 4.5 万~5 万 = 10, 5 万以上 = 11
X21	家庭是否拥有大病保险	是 = 1, 否 = 0	X22	家庭成员去年是否住院	是 = 1, 否 = 0
X23	家庭去年住院花费	0 = 0, 1 万以下 = 1, 1 万以上 = 2	X24	家庭近五年是否发生重大不良事件	是 = 1, 否 = 0
X25	家庭曾经是否为贫困户	是 = 1, 否 = 0	X26	家庭去年是否享有政府救助	是 = 1, 否 = 0
X27	家庭去年是否享有生产补贴	是 = 1, 否 = 0	Y	家庭是否潜在贫困风险户	是 = 1, 否 = 0

目前大多数研究以经济收入低于绝对贫困线标准为贫困依据,然而我国各地的扶贫标准大多是“两不愁三保障”,外加“一个收入”,“两不愁三保障”即不愁吃、不愁穿,保障其义务教育、基本医疗和住房[15]。然,要精准全面预测潜在返贫风险人群,应适当放宽标准,实现潜在返贫人群的早发现、早干预。基于数据库本身特征与研究所需,本研究以人均年收入低于 2019 年国家扶贫线 3218 元[19]、家庭享有住房保障、享有扶贫帮扶或风险边缘人群(低保户、监测户、低收入家庭)作为潜在贫困风险人群判定标准。

2.3. 统计分析

本研究利用 Spearman 相关性分析, 分析变量间的相关性, 剔除与因变量无相关性的变量, 自变量间相关系数绝对值大于 0.7 则保留其中一个变量, 变量筛选考虑在不改变模型拟合效果情况下, 数据尽可能简单化。研究变量属于高维变量, 利用 LASSO 回归对变量进行降维并筛选重要变量, 运用 Logistic 回归进行影响因素的探索, 并利用 Logistic 回归与 XGBoost 模型进行风险预测模型建模, 利用 ROC 曲线以及混淆矩阵综合评估模型预测效能, 最终选择 SHAP 模型对最优风险模型进行可视化处理。研究采用 stata17.0 对数据进行预处理, 采用 R4.4.1 对数据进行建模分析, 检验水准 $\alpha = 0.05$ 。

3. 研究结果

3.1. 相关性分析

将 28 个变量进行 Spearman 相关分析, 结果显示, “家庭人数”与“家庭是否贷款”跟家庭是否为潜在贫困风险户无相关性; “家庭成员去年是否住院”与“家庭去年住院花费”两变量存在高相关性, 相关系数为 0.96 ($P < 0.05$), 为保证模型尽可能简单, 故保留“家庭成员是否住院”纳入研究。见图 1。

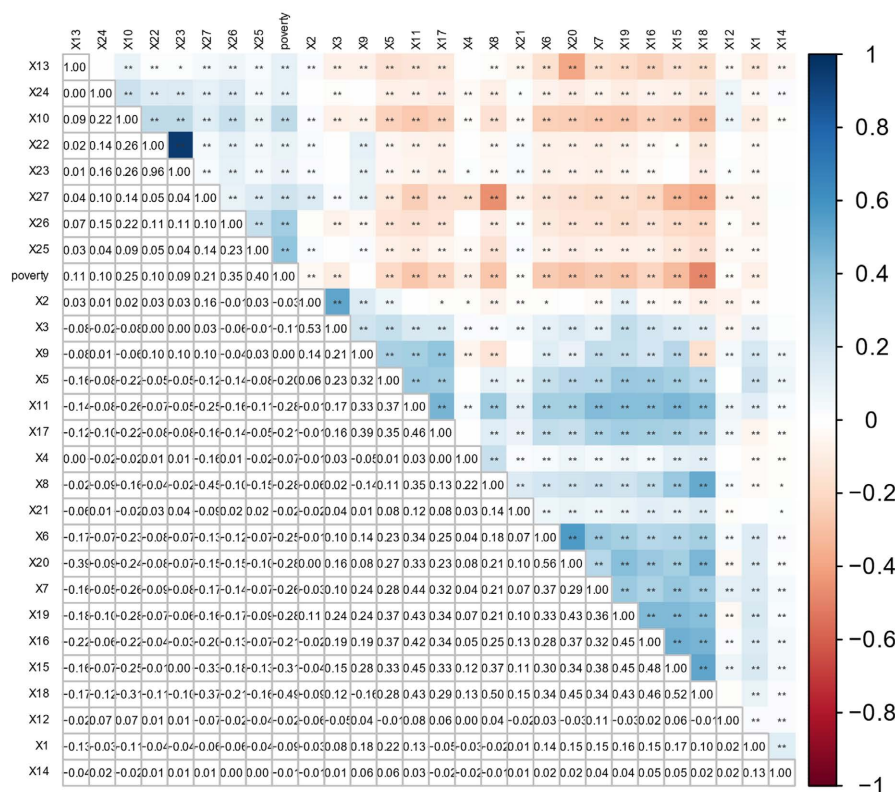


Figure 1. Variable correlation analysis

图 1. 变量相关性分析

3.2. LASSO 回归

由图 1 可见, 部分自变量间存在相关性, 且相关程度较高, 将剩余 24 个自变量纳入 LASSO 回归进行变量降维处理。结果显示, 随着 LASSO 回归的惩罚参数 λ 逐渐增大, 自变量系数逐渐压缩为 0 (见图

2(A)), 当 λ 取 $\lambda_{\min}$ 时, 模型筛选变量为 24 个, 此时模型二项偏差最小, 模型效能最优, 当参数 λ 取 λ_{1se} 时模型筛选变量 17 个, 此时模型性能较好且拥有变量最少(见图 2(B))。

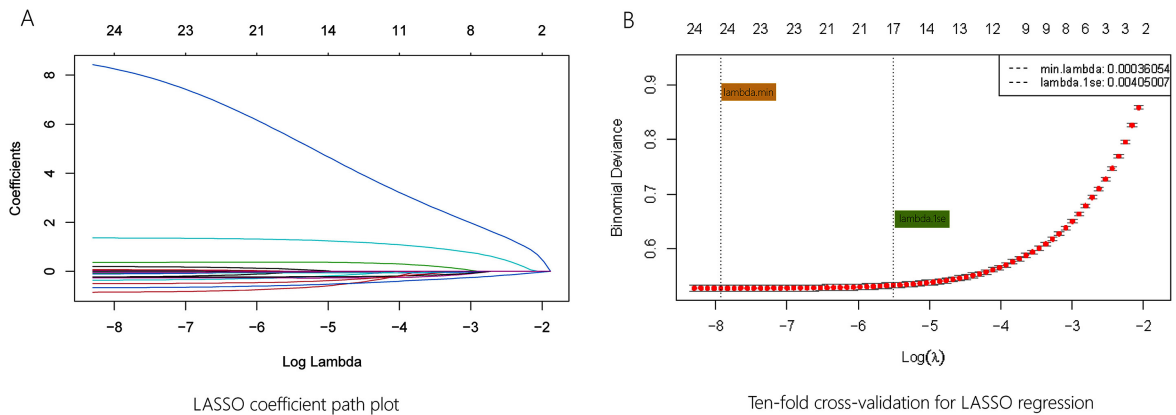


Figure 2. Variable selection based on LASSO regression
图 2. 基于 LASSO 回归的变量筛选

当 λ 取 λ_{1se} 时模型筛选表 2 所示工商活动、住房属性、住房数量等 17 个变量进行后续影响因素研究以及预测模型构建。

Table 2. LASSO variable selection table
表 2. LASSO 筛选变量表

变量	系数	变量	系数
工商活动	-0.042	住房属性	-0.688
住房数量	-0.012	汽车	-0.256
积蓄	-0.267	智能手机	-0.219
户口类型	-0.429	不健康人数比	0.382
学历	-0.052	非固定支出	0.003
工资性收入	-0.026	人均年收入	-0.567
耐用品价值	-0.052	存款	-0.031
大病保险	0.089	曾经贫困户	5.441
政府救助	1.289		

3.3. 多因素 Logistic 回归

以潜在返贫风险户为因变量, 将 LASSO 回归筛选的 17 个变量作为自变量纳入多因素 Logistic 回归模型, 选择逐步回归训练模型, 最终拟合模型排除变量“家庭住房数量”, 纳入其余 16 个变量, 模型似然比检验 $P < 0.001$, 伪 $R^2 = 0.554$, 表示模型拟合较好。表 3 所示, “家庭不健康人数比”、“家庭非固定支出”、“家庭拥有大病保险”、“曾经是贫困户”、“享有过政府帮扶”等因素为潜在返贫风险户的危险因素, 其余因素为保护因素。其中, 危险因素中, 曾经是否为贫困户影响最大, 曾经为贫困户的家庭潜在返贫风险是曾经非贫困户家庭的 8850.673 倍; 保护因素中, 家庭住房属性最为显著, 家庭住房为自有房的潜在返贫风险是非自有房家庭的 40.6%。

Table 3. Fitted parameters of the multivariate logistic regression model for households at risk of returning to poverty
表 3. 潜在返贫风险户的多因素 Logistic 回归模型拟合参数

变量	偏回归系数	标准误	Z	P	OR	95% CI
常数	2.084	0.125	16.670	<0.001	8.038	—
工商活动	-0.257	0.087	-2.946	0.003	0.773	(0.650, 0.916)
住房属性	-0.902	0.060	-15.072	<0.001	0.406	(0.361, 0.456)
汽车	-0.384	0.077	-5.016	<0.001	0.681	(0.585, 0.790)
积蓄	-0.249	0.050	-4.969	<0.001	0.780	(0.707, 0.860)
智能手机	-0.211	0.049	-4.284	<0.001	0.810	(0.735, 0.892)
户口类型	-0.531	0.057	-9.284	<0.001	0.588	(0.526, 0.658)
不健康人数比	0.389	0.064	6.069	<0.001	1.475	(1.301, 1.672)
学历(小学)	-0.235	0.091	-2.579	0.009	0.790	(0.662, 0.944)
学历(初中)	-0.343	0.091	-3.773	<0.001	0.710	(0.594, 0.848)
学历(高中)	-0.389	0.098	-3.984	<0.001	0.678	(0.559, 0.821)
学历(大学)	-0.230	0.107	-2.147	0.032	0.795	(0.644, 0.980)
学历(研究生及以上)	-0.197	0.332	-0.593	0.553	0.821	(0.411, 1.526)
非固定支出	0.076	0.013	5.713	<0.001	1.079	(1.051, 1.107)
工资性收入	-0.050	0.011	-4.595	<0.001	0.951	(0.930, 0.971)
人均年收入	-0.677	0.019	-34.790	<0.001	0.508	(0.489, 0.528)
耐用品价值	-0.068	0.019	-3.543	<0.001	0.935	(0.900, 0.970)
存款	-0.043	0.008	-5.181	<0.001	0.958	(0.943, 0.974)
大病保险	0.201	0.047	4.307	<0.001	1.223	(1.116, 1.340)
曾经贫困户	9.088	0.765	11.879	<0.001	8850.673	(2451.955, 5745.246)
政府救助	1.373	0.052	26.404	<0.001	3.949	(3.567, 4.373)

注：变量学历均以未上学作为参照；模型似然比检验 $P < 0.001$ ，Nagelkerke 伪 $R^2 = 0.554$ 。

采用十折交叉验证上述 Logistic 模型的预测效能，取 10 次验证数据的平均值评估模型预测效能，由图 3(A)可见模型的平均 ROC 曲线下面积 $AUC = 0.734$ ，由图 3(B)可计算得到模型精确率为 0.886、查准率为 0.785、灵敏度为 0.498、特异度为 0.970，F1 值为 0.609。

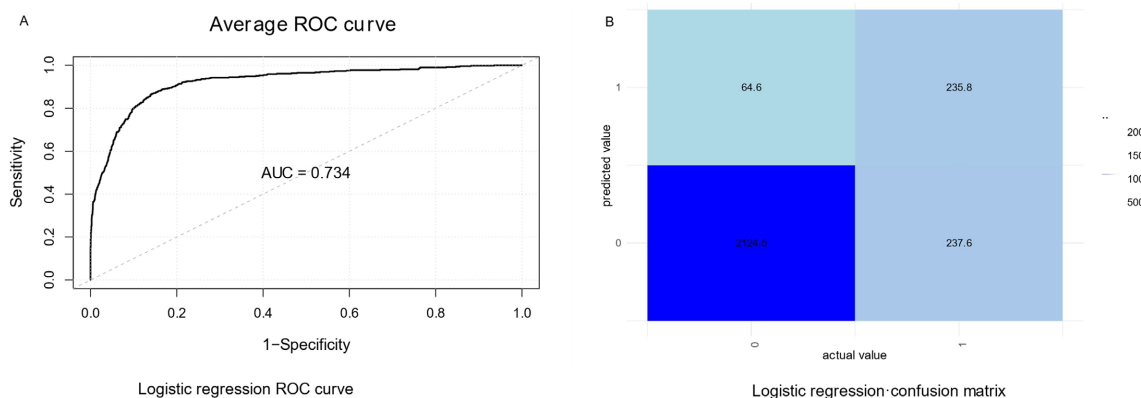


Figure 3. Logistic regression model predictive performance

图 3. Logistic 回归模型预测效能

3.4. XGBoost 预测模型

将多因素 Logistic 回归模型建模的 16 个变量纳入 XGBoost 模型。XGBoost 模型参数通过网格搜索法与十折交叉验证搜索最优组合，见表 4。通过十折交叉验证的平均结果用于评价模型的预测效能，由图 4(A)可见，模型 10 次训练的平均 ROC 曲线下面积 $AUC = 0.948$ ，由图 4(B)混淆矩阵可计算得到模型精确率为 0.905、查准率为 0.937、灵敏度为 0.949、特异度为 0.703，F1 值为 0.943。

Table 4. XGBoost model parameters
表 4. XGBoost 模型参数表

参数名	参数值	参数含义
max_depth	5	树的最大深度
eta	0.2	学习率
gamma	0.5	分裂最小损失
min_child_weight	2	最小子节点权重和
subsample	0.7	样本子采样率
colsample_bytree	0.4	特征子采样率
nround	60	迭代次数
objective	binary:logistic	二分类逻辑回归
metrics	rmse	均方根误差

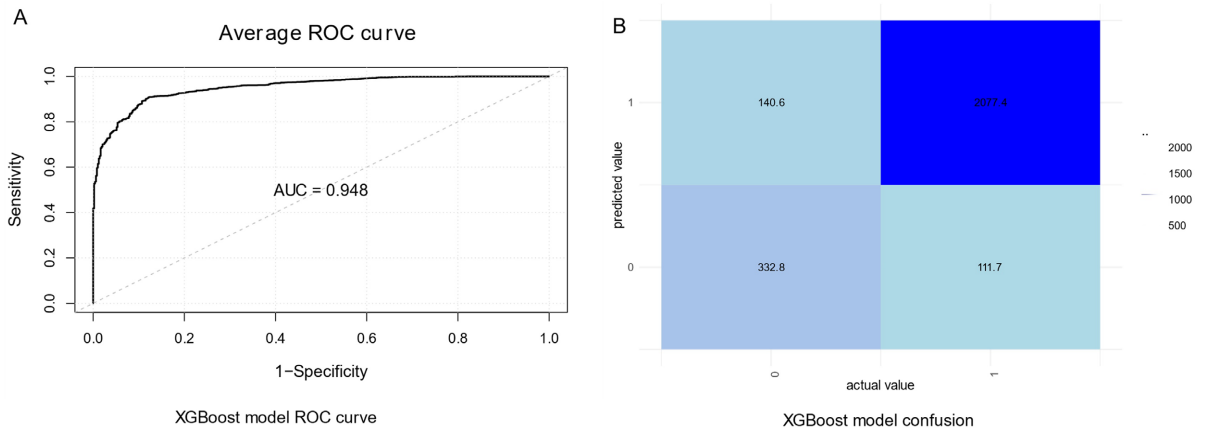


Figure 4. Predictive performance of the XGBoost model
图 4. XGBoost 模型预测效能

3.5. 模型评估

两模型采用受试者工作特征(receiver operating characteristic, ROC)曲线混淆矩阵的四个常用指标精确率(Accuracy)、查准率(Precision)、灵敏度(Sensitivity)、特异度(Specificity)，以及 F1 值对模型预测性能进行评估。由表 5 可知，Logistic 回归模型仅特异度高于 XGBoost 模型，其他评价指标均低于 XGBoost 模型。拟合模型往往追求灵敏度、特异度都比较高，本研究 Logistic 回归灵敏度远低于 XGBoost，那么 Logistic 回归模型就会有更高的漏诊率，本研究为了拟合潜在返贫风险模型，为全面尽早筛选高风险人群，故需要模型有更高的灵敏度。图 5 可见，XGBoost 模型的 ROC 曲线下面积显著高于 Logistic 回归($P < 0.05$)。

Table 5. Evaluation of the logistic regression and XGBoost models
表 5. Logistic 回归模型与 XGBoost 模型评估

指标	Logistic 回归	XGBoost 模型
精确率	0.886	0.905
查准率	0.785	0.937
灵敏度	0.498	0.949
特异度	0.970	0.703
F1 值	0.609	0.943

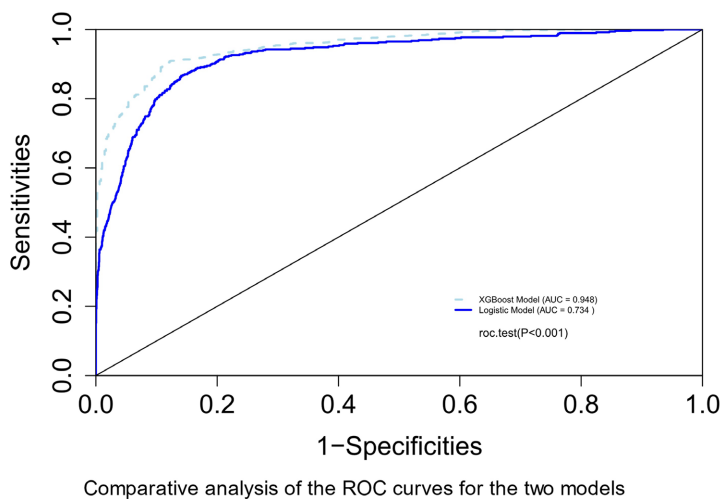


Figure 5. Comparison of the model’s ROC curves
图 5. 模型 ROC 曲线比较

3.6. 模型可视化

选用 XGBoost 模型作为潜在返贫风险户的预测模型，模型显示“家庭人均年收入”、“是否曾经是贫困户”、“是否享受政府帮扶”、“户主户口类型”、“家庭存款”是对模型贡献度最高的五个变量(图 6(A))。其中家庭人均年收入低于 5000 元潜在返贫风险最大，人均年收入大于 10,000 元，随着收入增加风险降低幅度变小；曾经是贫困户和享受有政府救助帮扶的家庭返贫风险较高，且对潜在返贫影响较大；农村户口相比城市潜在返贫风险更高；家庭存款 25,000 元以下，随着存款增加，潜在返贫风险降低，家庭存款高于 25,000 元，返贫风险有所波动，整体上随着存款增加，潜在返贫风险降低(图 6(B))。

4. 讨论

4.1. 潜在返贫风险的影响因素

由相关性结果可知，本研究中“家庭人数”与潜在返贫风险无关，而在其他研究既有发现家庭人数越多返贫风险越高[20]，又有数据显示家庭人数多会降低返贫风险[15]，家庭规模与潜在返贫风险的关系学术界尚未达成一致。“家庭住房数量”相关性分析显示与潜在返贫风险具有相关性，然而多因素 Logistic 回归模型却剔除该变量，可能变量“家庭住房属性”与之较为相关，相关系数为 0.53 ($P < 0.05$)。

多因素分析结果与现有研究较为一致，家庭不健康人数比越高、家庭非固定支出越高、家庭拥有大病保险、农村家庭、曾经是贫困户、享有过政府帮扶等家庭面临更高的潜在返贫风险，工资性收入越高、

拥有固定资产、积蓄越高的家庭潜在返贫风险越低[15][21]-[23]。家庭最高学历为小学、初中、高中相对于未上过学的家庭偏回归系数逐渐变小,即学历越高潜在返贫风险越低,这与非洲的一项研究一致[24],该研究发现中等教育能带来经济效益,且农村地区的经济增益高于城市;然而本研究发现家庭最高学历为大学的家庭,相比于未上过学的潜在返贫风险更低,但是偏回归系数有所增加,最高学历为研究生及以上的家庭与未上过学的家庭相比,潜在返贫风险差异无统计学意义($P=0.553$)。这可能是研究生及以上样本量太少的缘故;亦有可能是高等教育需要人力财力的投入,高等教育会增加学习年限,在一定时间范围内、一定程度上降低劳动力的经济创造能力,家庭返贫脆弱性较高的家庭会增加潜在返贫风险[25]。

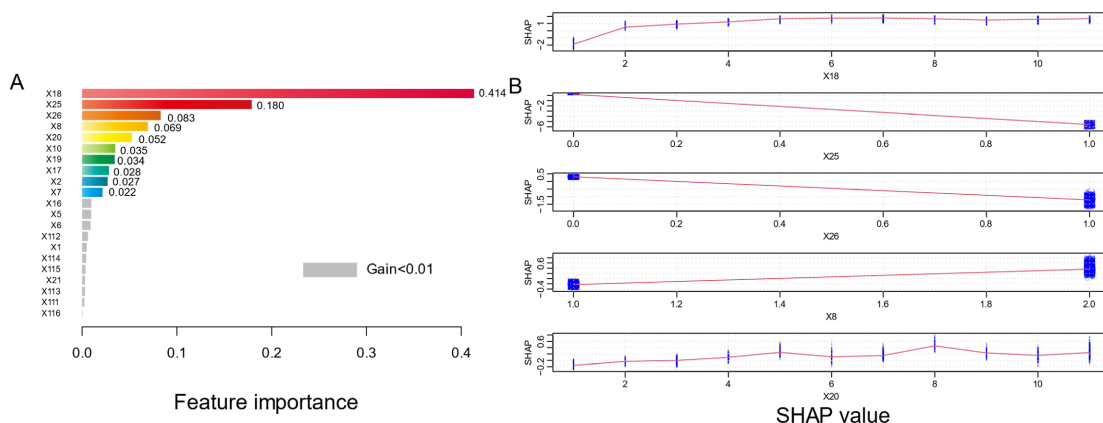


Figure 6. Visualization of the XGBoost model
图 6. XGBoost 模型可视化

4.2. 潜在返贫风险预测模型

Logistic 回归用于潜在返贫风险的预测,平均 AUC 为 0.734 这与 Logistic 回归预测其他领域效能较为一致[26]。本研究 Logistic 回归混淆矩阵和 ROC 曲线下面积均没有 XGBoost 表现优异,预测效能较 XGBoost 模型低,这与国内外在其他领域的研究结果大多一致[27]-[29]。而也有研究结果显示,XGBoost 模型与 Logistic 回归预测效能相似,且样本量较小 Logistic 回归更稳健[30]。故,机器学习并不一定完全优于传统统计方法,研究者亦需要结合数据类型选择合适的方法。

“家庭人均年收入”、“是否曾经是贫困户”、“是否享受政府帮扶”、“户主户口类型”、“家庭存款”是对 XGBoost 模型贡献度最高的五个变量,且“家庭人均年收入”的贡献度是其他变量的 2 倍以上这与 Logistic 回归的结果并非完全一致。XGBoost 模型可视化显示潜在返贫风险与家庭人均年收入在 1 万以下关系更为密切,可见 XGBoost 模型较 Logistic 回归能够更准确地发现变量内部变化的模型的影响。

4.3. 局限性分析

本研究数据为横断面数据,因故推断因果关系的能力受限,变量间的时间先后顺序及因果方向难以确定。其次,预测模型验证仅采用内部验证,模型泛化能力还有待进一步考量。此外,部分变量依赖自报数据,可能存在回忆偏倚或社会期许效应,影响数据准确性。下一步,研究可考虑采用纵向追踪数据,运用未来数据对模型进行外部验证,以增强模型的稳健性与解释力,从而为制定更精准的防返贫政策提供科学依据。

5. 结论与建议

从健康风险视角看,家庭成员健康状况,以及家庭拥有大病保险与潜在返贫风险息息相关。家庭成

员健康状况不好的家庭不仅有更大的医疗支出,同时因为疾病或伤害而无法工作,家庭的经济状况可能会更加困难[31][32];家庭拥有大病保险可能本身家庭成员患有或易患有重大疾病,相关政策制定者或研究人员应持续关注因病处于潜在返贫风险中的人群。

从贫困治理视角看,乡村振兴、教育脱贫是后扶贫时代贫困治理的必经之路[33]。已脱贫户与享有政府救助的家庭这些贫困边缘人群抵御返贫风险能力较差,更易面临潜在返贫风险,相关部门不仅要“授人与鱼”,更要“授人以渔”,全面推进乡村振兴,帮助贫困边缘人群,彻底摆脱贫困风险。科教兴国,教育能增加经济效益,相关部门应做好教育保障工作,打破“因贫辍学”、“因教返贫”的恶性循环,为降低潜在返贫风险提供内生动力。

基金项目

2026 年度重庆市科卫联合疾控科研项目,项目编号:2026KWJK1014 与 2027 年度重庆市科卫联合疾控科研项目(公共卫生人才项目),项目编号:2027KWJK1012 资助。

参考文献

- [1] World Bank (2020) Poverty and Shared Prosperity 2020: Reversals of Fortune. World Bank.
- [2] 2025 年分层分类社会救助体系不断完善[EB/OL].2026-01-02.
https://www.gov.cn/lianbo/202601/content_7053671.htm, 2026-07-07.
- [3] 姜晓萍, 郑时彦. 乡村振兴中规模性返贫风险的发生机理与阻断机制[J]. 理论与改革, 2023(1): 130-142+168.
- [4] 医保“三重保障”制度让我国因病返贫人数显著下降[EB/OL]. 2020-10-16.
http://www.xinhuanet.com/politics/2020-10/16/c_1126619621.htm, 2026-07-07.
- [5] 中国公益研究院. 新发展阶段社会政策[两会代表共议慈善医疗救助的统筹与联合][EB/OL]. 2021-03-04.
https://www.bnu1.org/show_2241.html, 2026-07-07.
- [6] Xu, L., Guo, M., Nicholas, S., Sun, L., Yang, F. and Wang, J. (2020) Disease Causing Poverty: Adapting the Onyx and Bullen Social Capital Measurement Tool for China. *BMC Public Health*, **20**, Article No. 63.
<https://doi.org/10.1186/s12889-020-8163-5>
- [7] 翟绍果. 健康贫困的协同治理: 逻辑、经验与路径[J]. 治理研究, 2018, 34(5): 53-60.
- [8] 刘艳丽, 刘政. 互联网嵌入、社会资本与农村居民贫困脆弱性[J]. 云南财经大学学报, 2023, 39(8): 85-100.
- [9] Jan, S., Laba, T., Essue, B.M., Gheorghe, A., Muhunthan, J., Engelgau, M., et al. (2018) Action to Address the Household Economic Burden of Non-Communicable Diseases. *The Lancet*, **391**, 2047-2058.
[https://doi.org/10.1016/s0140-6736\(18\)30323-4](https://doi.org/10.1016/s0140-6736(18)30323-4)
- [10] Jiang, Y., Huang, C., Yin, D., Liang, C. and Wang, Y. (2020) Constructing HLM to Examine Multi-Level Poverty-Contributing Factors of Farmer Households: Why and How? *PLOS ONE*, **15**, e0228032.
<https://doi.org/10.1371/journal.pone.0228032>
- [11] Sun, H., Li, X., Li, W. and Feng, J. (2022) Differences and Influencing Factors of Relative Poverty of Urban and Rural Residents in China Based on the Survey of 31 Provinces and Cities. *International Journal of Environmental Research and Public Health*, **19**, Article 9015. <https://doi.org/10.3390/ijerph19159015>
- [12] Bulthuis, S.E., Kok, M.C., Raven, J. and Dieleman, M.A. (2020) Factors Influencing the Scale-Up of Public Health Interventions in Low- and Middle-Income Countries: A Qualitative Systematic Literature Review. *Health Policy and Planning*, **35**, 219-234. <https://doi.org/10.1093/heapol/czz140>
- [13] Sunil, K., Ranjan, P.J. and Bahadur, D.R. (2023) Can Rural Livelihood Programs Enhance Capabilities and Reduce Vulnerability to Poverty? Evidence from a Tribal Region of Eastern India. *Economic Analysis and Policy*, **77**, 85-98.
- [14] 张学敏, 史玲燕, 薛艳, 等. 乡村振兴视阈下返贫预警评价指标体系构建与实证[J]. 统计与决策, 2021(13): 58-62.
- [15] 李泽峰, 左停, 林秋香. 推进乡村振兴过程中脱贫户防返贫预警工具构建——基于六个国家级脱贫县建档立卡数据的探索[J]. 农村经济, 2023(1): 53-62.
- [16] Liu, M., Hu, S., Ge, Y., Heuvelink, G.B.M., Ren, Z. and Huang, X. (2021) Using Multiple Linear Regression and Random Forests to Identify Spatial Poverty Determinants in Rural China. *Spatial Statistics*, **42**, Article 100461.
<https://doi.org/10.1016/j.spasta.2020.100461>

- [17] Grueso, H. (2023) Unveiling the Causal Mechanisms within Multidimensional Poverty. *Evaluation Review*, **47**, 1107-1134. <https://doi.org/10.1177/0193841x221140936>
- [18] 姚领, 谢家智. 贫困脆弱性的生成与测度: 基于风险“冲击-应对”的视角[J]. 统计研究, 2022, 39(11): 117-132.
- [19] 国务院扶贫办: 坚持“一个收入、两个不愁、三个保障”综合性脱贫标准不动摇[EB/OL]. 2020-03-12. <https://m.gmw.cn/baijia/2020-03/12/33642368.html>, 2026-07-07.
- [20] 钱鹏博, 程维朝, 江洋. 脱贫农村家庭人均收入影响与返贫防控分析——以安徽省亳州市为例[J]. 中国集体经济, 2023(1): 1-5.
- [21] 李壮壮, 龙莹. 新发展阶段农村多维贫困的识别与测度[J]. 华南农业大学学报(社会科学版), 2022, 21(2): 15-27.
- [22] Wang, B., Shi, H., Wang, Z., Xu, S., Deng, N., Qiu, Y., et al. (2023) Pandemics Erode Poverty Alleviation Process: Impact on Productive Livelihood and Poverty Return. *iScience*, **26**, Article 107177. <https://doi.org/10.1016/j.isci.2023.107177>
- [23] Lázár, A.N., Adams, H., Adger, W.N. and Nicholls, R.J. (2020) Modelling Household Well-Being and Poverty Trajectories: An Application to Coastal Bangladesh. *PLOS ONE*, **15**, e0238621. <https://doi.org/10.1371/journal.pone.0238621>
- [24] Werner, L.K., Ludwig, J., Sie, A., Bagagnan, C.H., Zabré, P., Vandormael, A., et al. (2022) Health and Economic Benefits of Secondary Education in the Context of Poverty: Evidence from Burkina Faso. *PLOS ONE*, **17**, e0270246. <https://doi.org/10.1371/journal.pone.0270246>
- [25] 王岚. 因教返贫与依教反贫: 乡村振兴背景下职业教育可持续性精准帮扶的困境与突破[J]. 黑龙江高教研究, 2024(3): 142-147.
- [26] Yan, Z., Wang, J., Dong, Q., Zhu, L., Lin, W. and Jiang, X. (2022) XGBoost Algorithm and Logistic Regression to Predict the Postoperative 5-Year Outcome in Patients with Glioma. *Annals of Translational Medicine*, **10**, 860. <https://doi.org/10.21037/atm-22-3384>
- [27] 颜兵倩. 基于 logistic 回归与 XGBoost 机器学习模型预测卒中相关性肺炎的对比研究[D]: [硕士学位论文]. 承德: 承德医学院, 2023.
- [28] 张利斌, 吴宗文. 基于 XGBoost 机器学习模型的信用评分卡与基于逻辑回归模型的对比[J]. 中南民族大学学报(自然科学版), 2023, 42(6): 846-852.
- [29] Ding, C., Guo, Y., Mo, Q. and Ma, J. (2022) Prediction Model of Postoperative Severe Hypocalcemia in Patients with Secondary Hyperparathyroidism Based on Logistic Regression and XGBoost Algorithm. *Computational and Mathematical Methods in Medicine*, **2022**, 1-7. <https://doi.org/10.1155/2022/8752826>
- [30] Larsson, A., Berg, J., Gellerfors, M. and Gerdin Wärnberg, M. (2021) The Advanced Machine Learner XGBoost Did Not Reduce Prehospital Trauma Mistriage Compared with Logistic Regression: A Simulation Study. *BMC Medical Informatics and Decision Making*, **21**, Article No. 192. <https://doi.org/10.1186/s12911-021-01558-y>
- [31] 崔严尹, 许子英, 蔡义国, 等. 健康冲击对中老年人家家庭致贫返贫的影响路径研究[J]. 卫生经济研究, 2023, 40(4): 43-47.
- [32] Ma, Y., Xiang, Q., Yan, C., Liao, H. and Wang, J. (2022) Poverty Vulnerability and Health Risk Action Path of Families of Rural Elderly with Chronic Diseases: Empirical Analysis of 1,852 Families in Central and Western China. *Frontiers in Public Health*, **10**, Article 776901. <https://doi.org/10.3389/fpubh.2022.776901>
- [33] 金华宝, 伍科. 乡村振兴促进共同富裕的三重逻辑[J]. 理论与改革, 2022(5): 69-81.