

机器学习预测中老年高血压风险的研究

冉娟*, 陈家旺, 殷爽, 李凌霄#

重庆理工大学数学科学学院, 重庆

收稿日期: 2026年6月23日; 录用日期: 2026年7月16日; 发布日期: 2026年7月27日

摘要

针对当前中老年高血压风险预测研究中模型泛化性不足、缺乏多方法对比与外部验证支撑等局限, 本文基于CHARLS中老年队列数据与NHANES独立外部验证数据, 构建“双重特征筛选-多模型对比-跨人群验证”的机器学习预测框架, 为中老年高血压风险分层筛查提供科学依据。该方法整合人口学、体格测量、血压指标、生活方式与慢性病变量, 采用单因素Logistic回归与Lasso回归联合筛选, 剔除多重共线性指标后确定13项核心变量; 构建Logistic回归、随机森林与XGBoost三种模型, 以ROC曲线、AUC等指标评估性能, 结合5折交叉验证与SHAP值解析验证模型稳健性与可解释性。结果表明, XGBoost模型综合性能最优, 内部AUC达0.942, 外部AUC为0.830, 内外AUC差值仅0.112, 无明显过拟合; 年龄、BMI、收缩压、舒张压为核心危险因素。本研究构建的XGBoost模型兼具最优区分效能与跨人群泛化稳定性, 可为中老年高血压早期筛查与个体化防控提供科学参考。此外, 基于非血压指标的补充分析模型同样展现出良好的早期预警效能, 与主模型形成分层互补, 为中老年高血压的社区初筛提供了无创化技术方案。

关键词

中老年高血压, 风险预测, 机器学习, XGBoost, 外部验证

Prediction of Hypertension Risk in Middle-Aged and Elderly Populations Using Machine Learning

Juan Ran*, Jiawang Chen, Shuang Yin, Lingxiao Li#

College of Science, Chongqing University of Technology, Chongqing

Received: June 23, 2026; accepted: July 16, 2026; published: July 27, 2026

*第一作者。

#通讯作者。

文章引用: 冉娟, 陈家旺, 殷爽, 李凌霄. 机器学习预测中老年高血压风险的研究[J]. 应用数学进展, 2026, 15(7): 263-278. DOI: 10.12677/aam.2026.157321

Abstract

This paper presents a machine learning prediction framework of “dual feature selection-multi-model comparison-cross-population validation” to address insufficient generalizability and lack of external validation in current hypertension risk prediction models for middle-aged and elderly populations. Based on CHARLS cohort data and independent NHANES validation data, we integrated multi-dimensional variables, such as demographic, anthropometric, blood pressure, lifestyle, and chronic disease variables, and performed joint feature selection using univariate Logistic regression and Lasso regression, identifying 13 core variables. Three models—Logistic Regression, Random Forest, and XGBoost—were constructed. Performance was evaluated using ROC curves and AUC, and robustness and interpretability were verified using 5-fold cross-validation and SHAP value analysis. Results showed that the XGBoost model exhibited the best overall performance, with an internal AUC of 0.942 and an external AUC of 0.830, a difference of only 0.112, indicating no significant overfitting. Age, BMI, systolic blood pressure, and diastolic blood pressure were identified as the core risk factors. The XGBoost model constructed in this study combines optimal discriminative power with cross-population generalization stability, providing a scientific reference for early screening and individualized prevention of hypertension in middle-aged and elderly populations. Furthermore, the supplementary analysis model based on non-blood pressure indicators also demonstrated good early warning efficacy, forming a hierarchical complement to the main model and providing a non-invasive technical solution for community-based initial screening of hypertension in middle-aged and elderly populations.

Keywords

Hypertension in Middle-Aged and Elderly Populations, Risk Prediction, Machine Learning, XGBoost, External Validation

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

随着人口老龄化进程加快,高血压已成为影响我国中老年人群健康的重大公共卫生问题。2023 年世界卫生组织《全球高血压报告》显示,高血压是全球致死致残的主要危险因素之一,近 30 年间患者数从 6.5 亿翻倍增至 13 亿,整体控制率仅为 21%,防控形势严峻[1]。《中国高血压防治指南(2024 年修订版)》数据表明,我国高血压患病人数接近 3 亿,中老年人群患病率超过 50%,但知晓率、治疗率、控制率仍处于较低水平[2]。高血压是诱发心脑血管疾病的首要危险因素,长期控制不佳将显著增加中老年人群不良心血管事件风险,高龄、糖尿病、血脂异常等关键危险因素在该群体中叠加,进一步加重疾病负担[3]。当前国家将慢性病防控与健康老龄化列为重点,提出推进慢性病早诊早治、提高基层数字化防控能力,相关专家建议也明确提出,需借助人工智能技术建立个体化风险预测模型,推动防控向“健康预防”转变[4]。

国内外学者围绕高血压风险预测开展了大量研究,Logistic 回归作为传统分析方法,可量化各因素与发病风险的关联强度[5];近年来,随机森林、XGBoost 等机器学习方法因能捕捉变量间非线性关系,被广泛用于构建预测模型,展现出良好效能[6]-[8]。研究表明,血压水平、BMI、血脂、年龄等均为高血压发病的重要影响因素,且不同年龄组病因构成存在差异[6][7]。然而,现有研究仍存在明显不足:传统统计模型难以捕捉危险因素间的非线性交互作用,高维数据拟合能力有限[5];机器学习模型普遍存在可解

释性差、易过拟合、外部验证不足等问题，泛化能力与临床实用性不强[6]；多数研究为单中心、回顾性设计，易产生选择偏倚[6]；预测指标多局限于传统静态变量，未充分挖掘常规体检指标的应用价值，难以满足基层大规模人群筛查的临床需求[2] [4]；同时，现有模型多面向普通人群，对中老年高危人群缺乏针对性设计，限制了其在基层医疗卫生机构的推广应用[5]。

鉴于我国中老年高血压管理仍以确诊后治疗为主，缺乏高危人群系统性风险预警，本研究基于中老年人群可及的常规体检指标，构建“双重特征筛选 - 多模型对比 - 跨人群验证”的机器学习预测框架，旨在挖掘中老年高血压发病关键危险因素，探究常规体检指标与高血压发病风险的关联，构建性能稳定、泛化性强的预测模型，为临床干预优化、基层防控策略制定提供科学依据，助力推进健康中国建设。

2. 数据来源及预处理

2.1. 数据与研究对象

本研究采用回顾性双队列设计，选取中国健康与养老追踪调查(CHARLS)标准化数据集作为建模数据，美国国家健康与营养检查调查(NHANES) 2017~2020 年数据库作为独立外部验证数据，以此检验模型跨人群泛化能力。两项数据库均采用科学抽样方法，具备人群代表性，涵盖人口学、体格指标、血压、生活方式及慢性病史等信息，数据质量可靠。

本研究研究对象设定为年龄 ≥ 45 岁中老年人，两组队列执行统一的纳入、排除标准与高血压诊断标准。纳入标准：年龄 ≥ 45 岁，人口学、体格及血压核心资料完整，可明确判定高血压患病状态。排除标准：年龄小于 45 岁、核心变量缺失、指标超出正常生理范围、二分类变量赋值异常样本。高血压诊断参照临床通用标准：收缩压 ≥ 140 mmHg 或舒张压 ≥ 90 mmHg，或既往确诊高血压并服用降压药物者判定为患病；血压正常、无高血压病史且未服药者判定为未患病。吸烟、饮酒、糖尿病、高血脂等指标均依据问卷报告结果统一判定。

本研究对 CHARLS 建模队列与 NHANES 外部验证队列全部纳入对象的人口学、体格临床、生活行为、慢性病基线资料进行组间对比，基线特征见表 1。

Table 1. Comparison of baseline characteristics between the CHARLS cohort and the NHANES cohort
表 1. CHARLS 队列与 NHANES 队列研究对象基线特征对比

特征	CHARLS (n = 18,942)	NHANES (n = 4233)	统计检验	P 值
人口学特征				
年龄(岁)	57.9 $\pm$ 9.7	62.5 $\pm$ 10.4	t 检验	<0.001
男性[n (%)]	9636 (50.9)	2086 (49.3)	卡方检验	0.064
身高(m)	1.59 $\pm$ 0.09	1.66 $\pm$ 0.10	t 检验	<0.001
体重(kg)	59.9 $\pm$ 11.3	79.9 $\pm$ 17.1	t 检验	<0.001
BMI (kg/m ²)	23.7 $\pm$ 3.6	28.9 $\pm$ 5.3	t 检验	<0.001
腰围(cm)	86.1 $\pm$ 10.1	100.8 $\pm$ 13.1	t 检验	<0.001
腰高比	0.54 $\pm$ 0.07	0.61 $\pm$ 0.08	t 检验	<0.001
临床特征				
收缩压(mmHg)	130.5 $\pm$ 21.1	130.6 $\pm$ 19.6	t 检验	0.798
舒张压(mmHg)	76.4 $\pm$ 12.0	75.2 $\pm$ 11.4	t 检验	<0.001
平均动脉压(mmHg)	94.4 $\pm$ 13.9	93.7 $\pm$ 12.6	t 检验	0.000
脉压(mmHg)	54.1 $\pm$ 15.4	55.4 $\pm$ 16.1	t 检验	<0.001

续表

糖尿病[n (%)]	1082 (5.7)	3336 (78.8)	卡方检验	<0.001
高血脂[n (%)]	1765 (9.3)	2146 (50.7)	卡方检验	<0.001
高血压[n (%)]	7595 (40.1)	2601 (61.4)	卡方检验	<0.001
行为学特征				
吸烟[n (%)]	7654 (40.4)	1358 (32.1)	卡方检验	<0.001
饮酒[n (%)]	7901 (41.7)	1355 (32.0)	卡方检验	<0.001

注：连续变量以 $\bar{x} \pm s$ 表示，组间比较采用独立样本 t 检验；分类变量以例数(百分比)表示，组间比较采用 χ^2 检验； $P < 0.05$ 代表组间差异具有统计学意义。

由表 1 基线对比可知，CHARLS 队列与 NHANES 外部验证队列在人口学、临床、生活行为及慢性病指标上大多存在统计学差异(P 均 <0.001)。CHARLS 受试者平均年龄 57.9 岁，较美国队列低 4.6 岁；其 BMI、腰围、腰高比均更低，整体肥胖程度更轻。慢性病方面，NHANES 人群糖尿病、高血脂、高血压患病率(78.8%、50.7%、61.4%)显著高于 CHARLS (5.7%、9.3%、40.1%)；国内队列吸烟、饮酒比例则更高。两组人群基线的全面异质性，可解释后续模型跨队列泛化能力的差异。

2.2. 变量体系与衍生指标

本研究选取两个数据集共有变量作为预测指标，结合临床意义与数据完整性，最终纳入人口学、体格测量、血压、生活方式、慢性病史五大类变量，并完成统一定义与赋值。为挖掘变量间非线性关联，提升模型预测效果，基于原始指标构建 6 项衍生特征，包括体质指数(BMI)、腰高比(WHtR)、平均动脉压(MAP)、脉压差(PP)、年龄平方、BMI 平方，各指标计算公式如下：

体质指数(BMI)：

$$BMI = \frac{weight}{height^2} \tag{1}$$

腰高比(WHtR)：

$$WHtR = \frac{waist}{height} \tag{2}$$

平均动脉压(MAP)：

$$MAP = DBP + \frac{SBP - DBP}{3} \tag{3}$$

脉压差(PP)：

$$PP = SBP - DBP \tag{4}$$

年龄平方(age_{sq})：

$$age_{sq} = age^2 \tag{5}$$

BMI 平方(bmi_{sq})：

$$bmi_{sq} = BMI^2 \tag{6}$$

2.3. 数据预处理流程

为降低数据缺陷对建模结果的干扰，本研究对两组数据集采用统一的标准化预处理流程，依次开展

缺失值处理、异常值剔除与数据集划分。缺失值处理采用分层策略：对高血压患病状态、年龄、血压等核心变量，若存在缺失则直接剔除对应样本；对疾病史、吸烟及饮酒等变量，结合问卷缺失机制分析，其缺失多由研究对象未主动报告所致，依据流行病学保守处理原则，采用常数填补法将缺失值统一赋值为 0(代表未患病、无相关行为)，在控制信息偏倚的同时避免样本量过度损失。在此基础上，依据临床生理正常范围与相关文献标准设定阈值，对身高、体重、血压、BMI 等连续变量进行异常值剔除，并剔除二分类变量取值非 0 或 1 的异常样本，以保证建模数据规范可靠。

数据划分采用三分验证方案：将 CHARLS 数据集按照 7:3 比例划分为训练集与内部测试集，分别用于模型训练、参数优化与内部效果评估；NHANES 数据集全程不参与建模，仅作为外部验证集，评价模型泛化性能。

3. 模型构建与评估

3.1. 特征筛选

为探究中老年高血压发病的关键危险因素，本研究采用单因素 Logistic 回归与 Lasso 回归联合筛选策略，剔除无关变量与多重共线性指标，最终确定核心建模变量。

3.1.1. Logistic 回归筛选

Logistic 回归是处理二分类因变量的标准非线性回归方法，模型表达式为：

$$p = \frac{1}{1 + e^{-(\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)}} \quad (7)$$

上式(7)中， P 为高血压患病概率， $X_1, X_2, \dots, X_p$ 为待筛选的影响因素， α 为常数项， $\beta_1, \beta_2, \dots, \beta_p$ 为回归系数。通过极大似然估计完成参数拟合，并采用逐步回归策略完成变量筛选，保留对结局具有独立影响的特征。

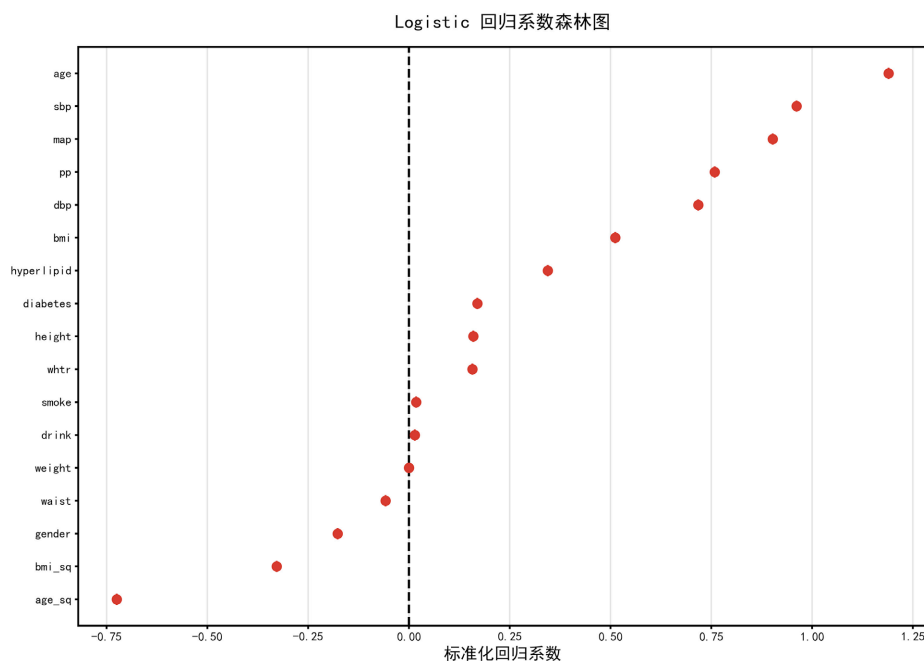


Figure 1. Forest plot of Logistic regression coefficients

图 1. Logistic 回归系数森林图

将预处理后的指标纳入模型，以回归系数及其显著性为依据筛选变量，并剔除多重共线性较强的指标。各变量的标准化回归系数及 95%置信区间如图 1 所示，结果显示年龄、收缩压、舒张压、体质指数、高血脂等为高血压的主要影响因素。

3.1.2. Lasso 回归筛选

Lasso 回归基于 L1 正则化实现特征筛选与系数压缩，目标函数为：

$$\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \tag{8}$$

上式(8)中， $\hat{\beta}$ 为回归系数估计值； y_i 为第 i 个样本的结局变量(高血压患病状态)； x_i 为第 i 个样本的自变量向量； β 为回归系数向量； λ 为正则化参数，用于控制系数压缩强度， λ 越大，被压缩至零的变量越多； n 为样本量； p 为自变量个数。

特征筛选过程如图 2 所示。各变量回归系数随惩罚系数 λ 变化的轨迹如图 2(a)所示，随着 λ 值增大，多数变量的系数逐渐被压缩至 0，直观呈现了变量的筛选过程。交叉验证下 λ 与错分误差的关系如图 2(b)所示，据此可确定最优 λ 值，实现模型拟合与特征精简的平衡，最终筛选出的 13 个变量回归系数结果如表 2 所示。

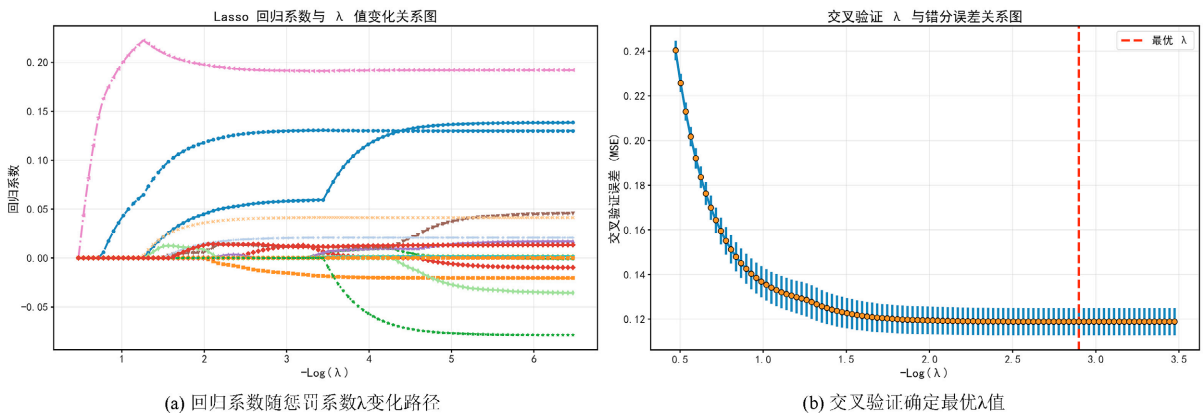


Figure 2. Lasso regression feature selection process
图 2. Lasso 回归特征筛选过程图

Table 2. Variables selected and their coefficients by Lasso regression
表 2. Lasso 回归筛选变量及系数

变量名	回归系数	变量名	回归系数
age	0.1175	drink	0.0017
gender	-0.0189	diabetes	0.0209
height	0.0091	hyperlipid	0.0414
weight	0.0007	map	0.0024
bmi	0.0224	agesq	-0.0576
waist	0.0103	sbp	0.2567
dbp	0.0739		

3.2. 模型构建与性能对比

基于 Lasso 回归筛选得到的 13 项核心特征，分别构建逻辑回归、随机森林、XGBoost 三种二分类预

测模型，采用分层交叉验证完成参数寻优，从混淆矩阵、ROC 曲线及多项效能指标开展模型评估与对比，筛选最优预测模型。

3.2.1. 模型构建

逻辑回归：以筛选特征为输入，高血压患病状态为结局变量构建模型，设置类别权重平衡样本分布。数据集按 7:3 分层划分为训练集与内部测试集，统一采用训练集参数完成特征标准化。结合交叉验证与约登指数确定最优分类阈值，作为模型性能参照基线。

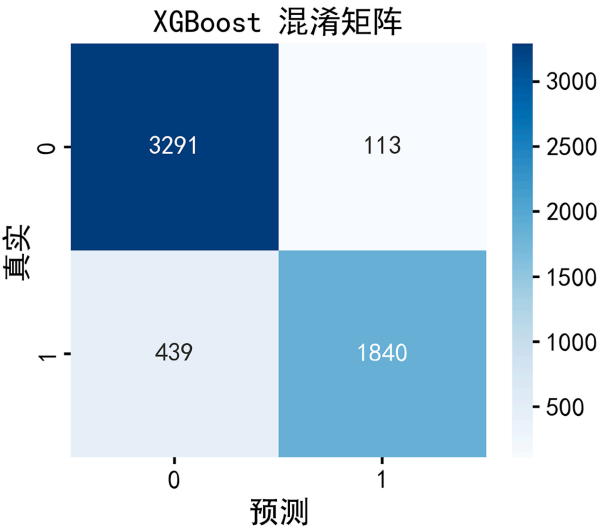
随机森林：数据划分、预处理流程与逻辑回归保持一致。采用网格搜索联合 5 折分层交叉验证，以交叉验证 AUC 最大为目标优化树数量、树深度等参数，完成模型训练后，在内部测试集与外部验证集评估性能。

XGBoost：沿用相同数据处理规则，引入样本权重缓解类别不平衡。通过网格搜索与 5 折分层交叉验证，对学习率、正则系数等参数联合寻优，在内外数据集上检验模型区分能力与泛化效果。

超参数优化：三种模型均采用分层 5 折交叉验证结合网格搜索的寻优策略，以验证集 AUC 为优化目标，针对不同算法设定专属参数区间。确定最优参数后在完整训练集重新拟合模型，降低数据随机划分带来的偏差。

3.2.2. 模型评价与结果分析

混淆矩阵：混淆矩阵可直观反映模型分类情况，是计算灵敏度、特异度的基础。三种模型在内部测试集的混淆矩阵见表 3~5，最优 XGBoost 模型分类结果以热力图展示。



注：图中 0 代表非高血压人群，1 代表高血压人群；横轴是模型的预测结果，纵轴为实际标签。

Figure 3. Confusion matrix of the optimal XGBoost model
图 3. 最优 XGBoost 模型混淆矩阵

Table 3. Confusion matrix of the Logistic regression model on the internal test set
表 3. 逻辑回归模型在内部测试集的混淆矩阵

真实类别\预测类别	0 (非高血压)	1 (高血压)
0 (非高血压)	3104	300
1 (高血压)	419	1860

Table 4. Confusion matrix of the random forest model on the internal test set
表 4. 随机森林模型在内部测试集的混淆矩阵

真实类别\预测类别	0 (非高血压)	1 (高血压)
0 (非高血压)	3358	46
1 (高血压)	517	1762

Table 5. Confusion matrix of the XGBoost model on the internal test set
表 5. XGBoost 模型在内部测试集的混淆矩阵

真实类别\预测类别	0 (非高血压)	1 (高血压)
0 (非高血压)	3290	114
1 (高血压)	441	1838

结合表 3~5 与图 3，XGBoost 模型的真阴性与真阳性样本数最高、假阳性与假阴性样本数最低，对高血压患者与非患者的区分能力最优，为后续效能评估提供了可靠依据。

ROC 曲线分析：ROC 曲线及曲线下面积 AUC 用于评价模型整体区分能力，AUC 越接近 1，模型性能越好。三种模型 ROC 曲线对比结果见图 4。

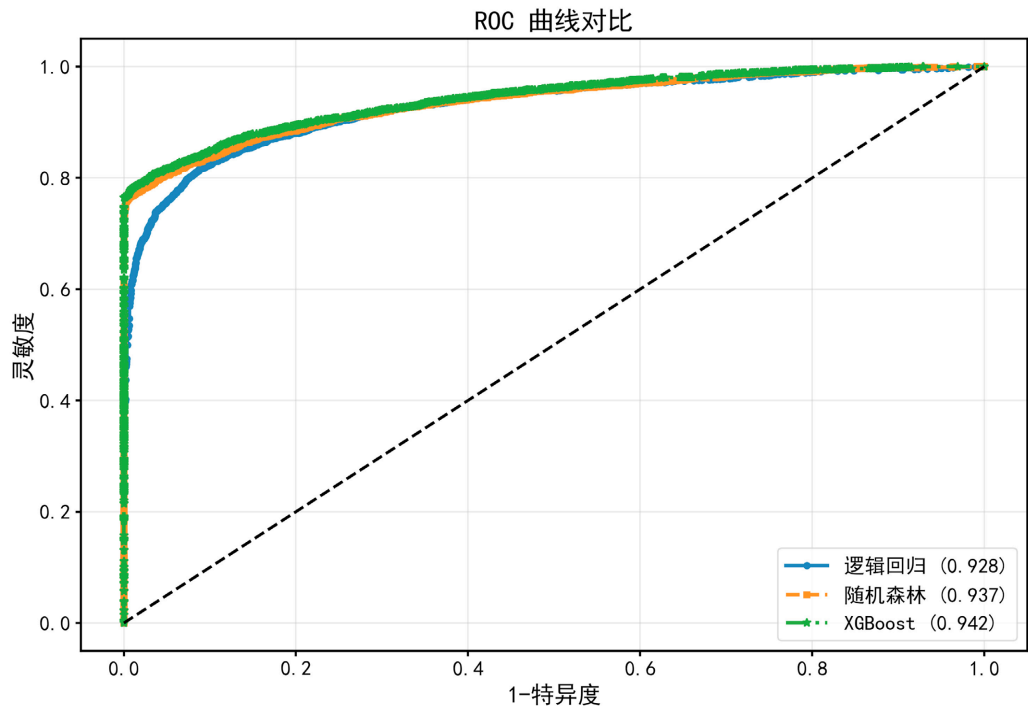
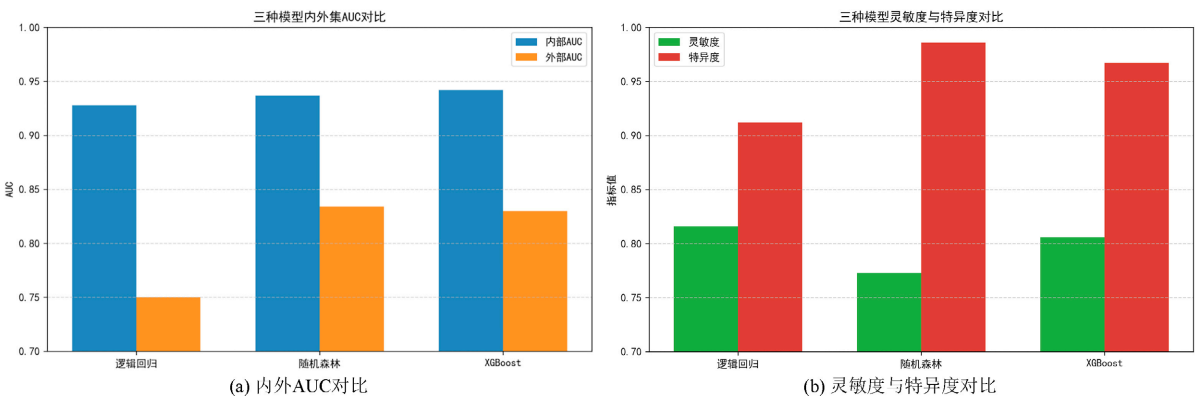


Figure 4. Comparison of ROC curves for three models
图 4. 三种模型 ROC 曲线对比

结果显示，XGBoost、随机森林、逻辑回归的 AUC 分别为 0.942、0.937、0.928。集成学习模型区分能力优于传统线性模型，该结论与混淆矩阵分析结果一致。

综合性能对比：为直观呈现各模型泛化稳定性与筛查效能差异，分别绘制内外 AUC 对比柱状图与灵敏度、特异度对比柱状图，结果见图 5(a)、图 5(b)。



注：图(a)为内部测试集与外部验证集 AUC 值对比，AUC 差值越小提示模型泛化稳定性越强；图(b)为模型灵敏度与特异度对比，分别反映模型对高血压病例的识别能力与对健康人群的排除能力。

Figure 5. Performance indicators comparison of three models

图 5. 三种模型性能指标对比图

Table 6. Performance comparison results of three models

表 6. 三种模型性能对比结果

模型	内部 AUC	外部 AUC	灵敏度	特异度
逻辑回归	0.928	0.750	0.816	0.912
随机森林	0.937	0.834	0.773	0.986
XGBoost	0.942	0.830	0.806	0.967

注：AUC 差值 = 内部 AUC - 外部 AUC，差值越小，模型泛化稳定性越强。

综合表 6、ROC 曲线及图 5 的对比结果，逻辑回归模型内外 AUC 差值较大，跨数据集泛化稳定性有限，但灵敏度表现最优，对高血压病例的识别能力较强；集成学习模型整体区分效能与泛化性能均优于传统线性模型，其中 XGBoost 模型内部 AUC 最高，且内外 AUC 差值较小，同时在灵敏度与特异度之间实现了最佳平衡，综合性能最优。

为进一步明确非血压指标的早期筛查贡献，本研究额外构建了剔除 SBP、DBP、MAP 及 PP 的 XGBoost 辅助模型。该模型内部 AUC 为 0.697，外部 AUC 为 0.671，虽低于主模型，但作为完全基于非血压信息的“前置预警”工具，其区分能力仍处于临床可接受范围，证实了常规非血压体检指标在高血压风险分层中的独立应用价值。

3.3. 特征解释与风险因素分析

为进一步揭示模型预测机制、明确中老年高血压患病的关键危险因素，本研究采用模型内置特征重要性与 SHAP 值两种方法开展模型可解释性分析，从不同维度解析变量的作用强度与影响机制。

3.3.1. 模型内置特征重要性分析

随机森林与 XGBoost 模型均可通过内置特征重要性指标，量化各输入变量对模型预测结果的贡献大小。特征重要性得分越高，表明该变量在模型决策过程中占据的权重越大，对高血压患病风险预测的影响越显著。

分别提取两种模型的特征重要性得分，按重要性由高到低排序并绘制柱状图，结果见图 6(a)、图 6(b)。该结果直观呈现了高血压预测的核心影响因素，结合临床意义对关键变量进行解读，可为人群高血压风

险筛查与早期干预提供参考。本研究得到的核心危险因素排序与现有研究结论基本一致[9][10]。

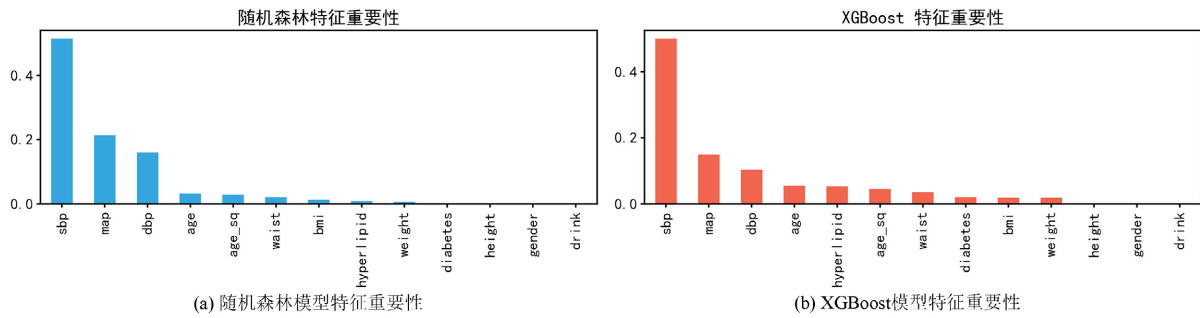


Figure 6. Comparison of feature importance between two models
图 6. 两种模型特征重要性对比

3.3.2. 基于 SHAP 值的模型解释

SHAP 值基于博弈论框架，能够对模型预测结果进行归因解释，量化各特征对模型输出的影响方向与作用强度，是当前机器学习模型可解释性研究中公认的有效方法[11]。本研究采用 TreeExplainer 模块对最优 XGBoost 模型计算 SHAP 值，开展模型归因解释分析。

与模型内置的特征重要性不同，SHAP 值不仅能体现特征的重要性排序，还可揭示其影响方向与作用机制。为直观展示各特征对模型预测结果的整体贡献程度，基于各变量的平均绝对 SHAP 值降序绘制特征贡献度条形图，结果见图 7。

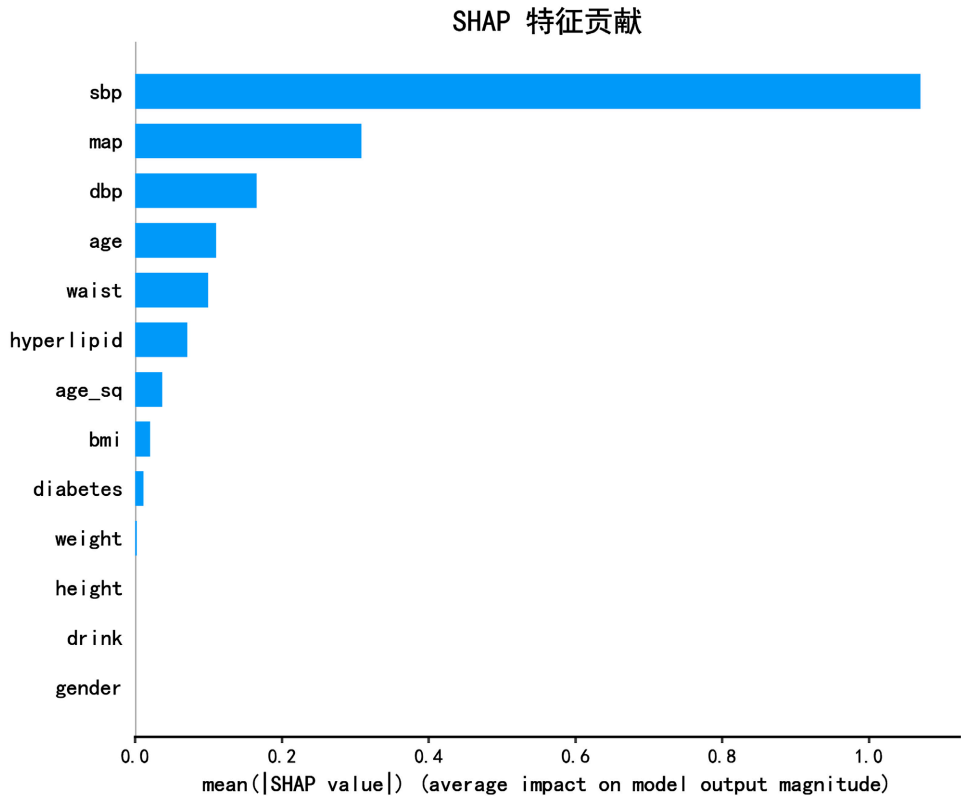


Figure 7. SHAP feature contribution of the XGBoost model
图 7. XGBoost 模型 SHAP 特征贡献度

分析结果显示,收缩压(sbp)的贡献度最高,是影响模型预测结果的核心变量;其次为平均动脉压(map)、舒张压(dbp)与年龄(age),与临床认知高度一致。SHAP 分析结果与随机森林、XGBoost 的特征重要性结果相互印证,共同明确了高血压发生的核心危险因素,为模型提供了解释依据,有效提升了模型的临床接受度与应用价值。

3.4. 模型稳健性分析

3.4.1. 偏倚风险评估

本研究通过分层抽样划分数据集,保证训练集与测试集样本分布一致,降低选择偏倚。同时通过 Lasso 回归筛选变量,规避多重共线性问题,减少混杂因素干扰。

为验证建模数据质量,绘制核心指标分布直方图,结果见图 8。

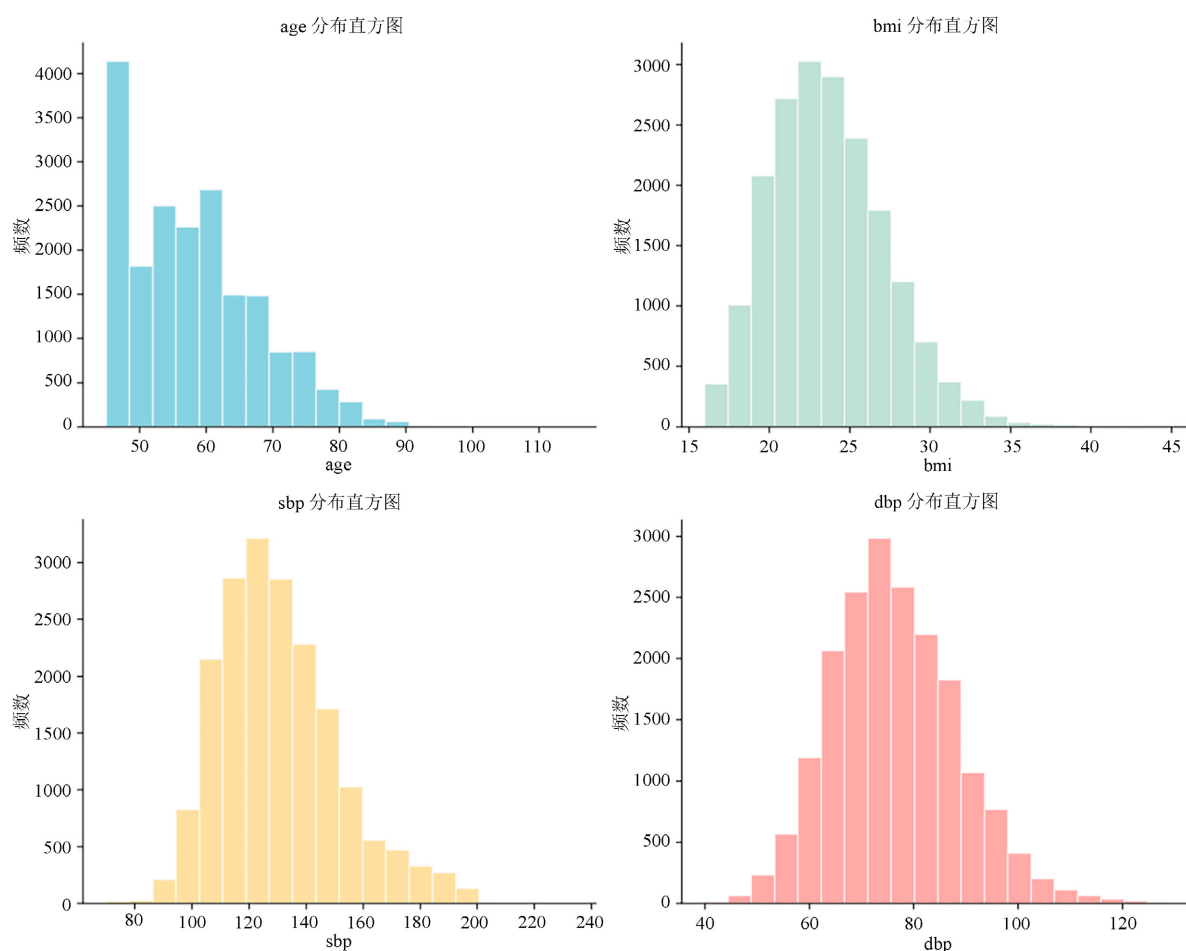


Figure 8. Histograms of core modeling indicators

图 8. 核心建模指标分布直方图

由图可知,各变量无明显极端异常值与严重偏态分布,数据整体质量良好,满足建模要求。

进一步绘制变量相关性热力图检验共线性,结果见图 9。

绝大多数变量相关系数低于 0.7,仅部分血压指标存在中度相关,无严重共线性问题,变量独立性较好。内外数据集的性能差异属于人群异质性导致的客观偏倚,模型无系统性偏倚。

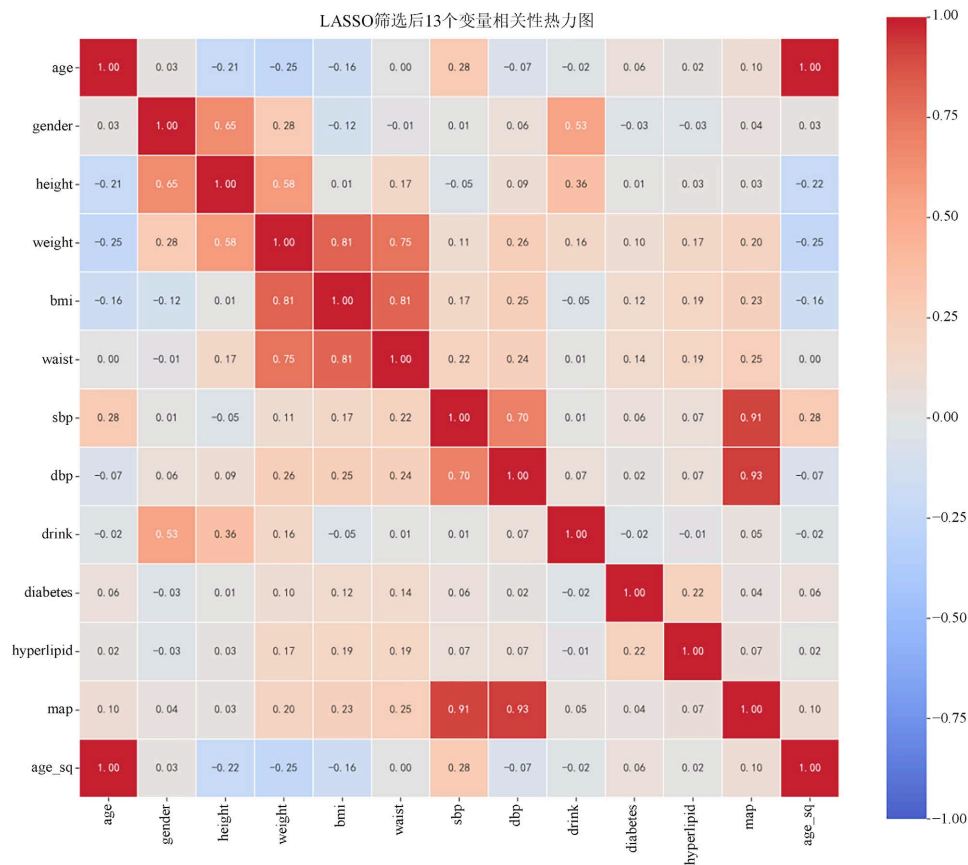


Figure 9. Correlation heatmap of modeling variables
图 9. 建模变量相关性热力图

3.4.2. 稳健性检验

为验证模型结果的稳定性，本研究分别开展缺失值敏感性分析与交叉验证分析。

缺失值敏感性分析：本研究各指标缺失比例均低于 5%。分别采用删除法与多重插补法处理缺失值并构建模型，两种方式所得模型各项评价指标差异极小，表明缺失值处理方式对模型结果影响有限，模型稳健性良好。

交叉验证分析：采用 5 折分层交叉验证对模型进行重复训练，各折训练集与验证集 AUC 结果如表 7 所示，波动趋势见图 10。

Table 7. AUC results of 5-fold cross-validation
表 7. 5 折交叉验证 AUC 结果

折数	训练集 AUC	验证集 AUC
1	0.946	0.950
2	0.947	0.948
3	0.949	0.937
4	0.946	0.947
5	0.949	0.937
均值	0.947	0.944

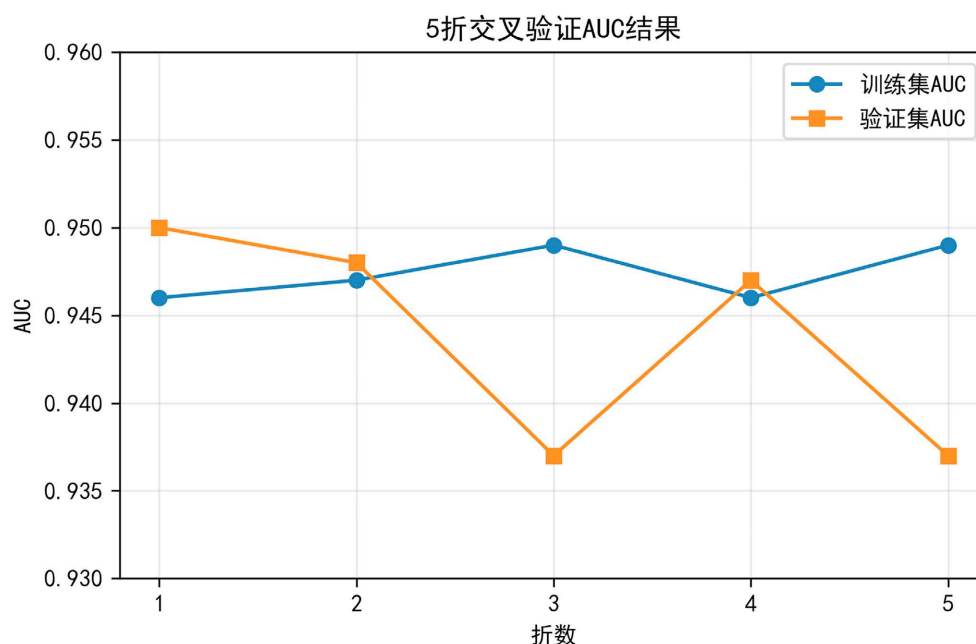


Figure 10. AUC line chart of 5-fold cross-validation

图 10. 5 折交叉验证 AUC 折线图

结果显示，训练集与验证集平均 AUC 差值仅为 0.003，各折性能波动较小，模型在不同数据子集下表现稳定，无明显过拟合，可重复性良好。

3.4.3. 两队列基线异质性的对模型跨区域泛化性能的影响

结合表 1 基线对比结果，CHARLS 与 NHANES 队列人群基线存在多维度统计学差异，直接造成模型跨队列预测效能衰减。在人口体型层面，NHANES 人群平均年龄、BMI、腰围、体重均显著高于 CHARLS 人群，人种躯体特征差异使得 BMI、腰围等指标对高血压的风险贡献权重存在人群特异性；慢性病层面，NHANES 高血压患病率 61.4%，远高于 CHARLS 的 40.1%，糖尿病、高血脂患病率差距更为悬殊，患病分布偏移会干扰模型概率校准与统一风险阈值的适用性；行为习惯层面，CHARLS 中老年吸烟、饮酒占比更高，中美人群高盐、高脂饮食、肥胖暴露模式不同，进一步改变各危险因素的预测贡献；同时两队列血压测量流程、诊断细则存在细微区别，血压基线指标分布差异进一步放大跨人群预测偏差。

人群基线异质性是模型外部验证 AUC 下降的核心诱因：一是东亚与欧美人群遗传易感特征不同，危险因素效应存在种族差异；二是饮食、肥胖等生活暴露模式差异改变变量预测权重；三是两组人群高血压基线患病率差距大，破坏模型概率校准效果；四是血压测量、临床诊断标准不完全统一，增加跨队列预测偏倚。

3.4.4. 模型局限性

模型在内部数据集中拟合效果稳定，但外部验证 AUC 出现一定下降，主要受不同人群地域、生活方式及疾病特征差异影响，跨区域推广仍需进一步校准。同时，本研究基于约登指数确定单一截断值，仅适用于本研究数据场景，难以适配全部临床筛查需求。

3.5. 补充分析：基于非血压指标的早期风险预测模型

为评估常规非血压体检指标在高血压发生前的独立预警价值，本研究在剔除收缩压(SBP)、舒张压(DBP)、平均动脉压(MAP)和脉压(PP)共 4 项血压相关指标后，保留年龄、性别、身高、体重、BMI、腰

围、腰高比(WHtR)、吸烟、饮酒、糖尿病、高血脂、年龄平方及 BMI 平方共 13 项非血压变量，重新构建预测模型。数据预处理、样本划分(7:3)、5 折分层交叉验证及超参数网格搜索策略均与主模型完全一致，算法选用综合性能最优的 XGBoost。

该辅助模型在内部测试集上的 AUC 为 0.697，在 NHANES 外部验证集上的 AUC 为 0.671。虽然其预测精度显著低于纳入血压指标的主模型(内部 AUC 0.942，外部 AUC 0.830)，但作为完全脱离血压读数的“前置筛查”工具，其仍展现出良好的风险分层能力。结果表明，年龄增长、超重、肥胖、不良生活方式及慢性病史等非血压因素已构成高血压发病的重要风险基底，即便在无法获取血压值的场景下，依托常规体格检查与问卷信息仍可有效甄别高危个体。

该辅助模型与主模型形成明确的互补应用格局：主模型适用于具备标准化血压测量条件的医院和体检中心，用于精准辅助诊断；非血压辅助模型则更契合基层社区卫生站大规模流行病学调查、线上自评小程序以及前瞻性队列研究中的早期风险预警，具备更广的人群覆盖面与推广便捷性。

即便在无法获取血压值的场景下，依托常规体格检查与问卷信息仍可有效甄别高危个体。为直观对比主模型与非血压辅助模型的区分效能，汇总两类模型内、外部验证的各项评价指标，模型性能对比结果详见表 8。

Table 8. Performance comparison between the primary model and the supplementary model
表 8. 主模型与辅助模型性能对比

模型	内部 AUC	外部 AUC	灵敏度	特异度
XGBoost 主模型(含血压指标)	0.942	0.830	0.806	0.967
XGBoost 辅助模型(非血压指标)	0.697	0.671	0.615	0.668

3.6. 风险分层与临床应用

本研究以模型输出的高血压患病预测概率为依据，结合临床筛查与分级管理实际需求，确定 $P_1 = 0.15$ 及 $P_2 = 0.42$ 两级风险阈值，据此将研究对象划分为低、中、高三个风险等级，分层标准参照《中国高血压防治指南(2024 年修订版)》及 2018 ESC/ESH 指南相关推荐[12] [13]。低风险人群患病概率极低，无明显危险因素聚集；中风险人群存在一定发病风险，且伴随部分可干预危险因素；高风险人群危险因素暴露水平较高，高血压发病风险显著升高，可为分层干预提供依据。

依据国内外高血压防控指南相关要求[12] [13]，针对不同风险等级人群制定差异化分级干预策略，并构建轻量化风险评估工具。低风险人群以健康维持与一级预防为重点，中风险人群以危险因素管控与风险逆转为核心，高风险人群采取强化干预与临床转诊相结合的管理模式。同时基于特征重要性及 SHAP 值分析结果构建简易风险评分卡，通过信息录入、风险评分、等级判定与干预推荐的标准化流程，实现快速风险评估与个性化健康指导，具备良好的基层推广价值。

4. 总结

本研究基于 CHARLS 中老年人队列数据，构建并验证了逻辑回归、随机森林与 XGBoost 三种高血压患病预测模型，结果显示 XGBoost 模型综合性能最优，其内部测试集 AUC 为 0.942，外部独立验证 AUC 为 0.830，泛化稳定性良好；通过特征重要性及 SHAP 值分析明确了年龄、血压、BMI 等为高血压发病的核心危险因素，并基于模型预测概率构建了三级风险分层体系，为人群精细化防控提供了依据。本研究采用双队列研究设计，结合约登指数法确定最优截断值，形成了“预测模型 - 风险分层 - 分级干预 - 轻量化工具”的完整研究闭环，同时，补充构建的非血压指标辅助模型有效拓展了预测工具的应用

边界,在缺乏血压测量条件的场景下仍可实现早期风险预警,形成了“精准诊断”与“广泛筛查”双轨并行的技术格局,提升了研究成果向基层转化的适配性,为基层高血压防控提供了可落地的技术方案[14]-[17]。然而,受不同人群地域与生活方式差异影响,模型跨队列验证时性能存在一定衰减,且外部验证数据来源单一、未纳入多维度临床特征,模型普适性与预测效能仍有提升空间[18][19]。

针对上述人群基线异质性带来的模型泛化衰减问题,为提高模型的跨区域适用性,建议采取以下校准策略:① 对 BMI、腰围等体型指标进行人群特异性标准化处理;② 在目标人群数据上进行模型微调,重新校准截距和系数;③ 采用患病率加权方法调整预测概率;④ 按年龄、性别等分层评估模型性能,识别性能较差的亚人群并针对性优化;⑤ 考虑加入人群特异性特征或交互项,以捕捉不同人群中危险因素的效应差异。未来可通过纳入多中心数据、整合血生化及影像学指标、开展前瞻性验证等方式进一步优化模型,推动其在真实世界高血压防控中的临床转化应用。

基金项目

重庆市大学生创新创业训练计划项目(2025CXXL203),重庆市教委科学技术研究项目(青年),项目号:KJQN202401139。

参考文献

- [1] 张冬燕,李燕.世界卫生组织《全球高血压报告》(2023年)概要及解读[J].诊断学理论与实践,2024,23(3):297-304.
- [2] 刘嘉慧,刘靖.《中国高血压防治指南(2024年修订版)》亮点、要点解读[J].中国医学前沿杂志(电子版),2025,17(1):1-6.
- [3] 钱晰彦,汤一帆,李婷茹,等.高血压患者发生心血管疾病相关危险因素的 Meta 分析[J].中国循证心血管医学杂志,2024,16(12):1416-1423.
- [4] 中国高血压联盟《高血压患者高质量血压管理中国专家建议》委员会.高血压患者高质量血压管理中国专家建议[J].中华高血压杂志(中英文),2024,32(2):104-111.
- [5] 曲帅,张盈,陈雪,等.高血压患病风险预测模型的研究进展[J].心脏杂志,2025,37(2):216-220+235.
- [6] 魏诗意,张珍,王叶婷,等.基于机器学习构建高血压患病风险预测模型的系统评价[J].牡丹江医学院学报,2024,45(5):55-61.
- [7] 沈赛拉,钟锋,梁兴,等.基于随机森林和梯度提升决策树的高血压分析预测[J].计算机时代,2023(5):15-19.
- [8] 张添玉,邢春国,马雨杨,等.基于深度学习的高血压风险预测模型的构建[J].医药高职教育与现代护理,2026,9(2):172-178.
- [9] Xu, T., Liu, J., Liu, J., et al. (2021) Risk Factors for Hypertension in Chinese Adults: A Systematic Review and Meta-Analysis. *Journal of Hypertension*, **39**, 1110-1119.
- [10] Zhao, D., Liu, J., Wang, M., Zhang, X. and Zhou, M. (2019) Epidemiology of Cardiovascular Disease in China: Current Features and Implications. *Nature Reviews Cardiology*, **16**, 203-212. <https://doi.org/10.1038/s41569-018-0119-4>
- [11] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, 4-9 December 2017, 4765-4774.
- [12] 中国高血压防治指南修订委员会.中国高血压防治指南(2024年修订版)[J].中国心血管杂志,2024,29(1):1-56.
- [13] Williams, B., Mancia, G., Spiering, W., Agabiti Rosei, E., Azizi, M., Burnier, M., et al. (2018) 2018 ESC/ESH Guidelines for the Management of Arterial Hypertension. *European Heart Journal*, **39**, 3021-3104. <https://doi.org/10.1093/eurheartj/ehy339>
- [14] Riley, R.D., Ensor, J., Snell, K.I.E., Debray, T.P.A., Altman, D.G., Moons, K.G.M., et al. (2016) External Validation of Clinical Prediction Models Using Big Datasets from E-Health Records or IPD Meta-Analysis: Opportunities and Challenges. *BMJ*, **353**, i3140. <https://doi.org/10.1136/bmj.i3140>
- [15] Youden, W.J. (1950) Index for Rating Diagnostic Tests. *Cancer*, **3**, 32-35. [https://doi.org/10.1002/1097-0142\(1950\)3:1<32::aid-cnrcr2820030106>3.0.co;2-3](https://doi.org/10.1002/1097-0142(1950)3:1<32::aid-cnrcr2820030106>3.0.co;2-3)
- [16] Topol, E.J. (2019) High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature*

- Medicine*, **25**, 44-56. <https://doi.org/10.1038/s41591-018-0300-7>
- [17] Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V.X., Doshi-Velez, F., *et al.* (2019) Do No Harm: A Roadmap for Responsible Machine Learning for Health Care. *Nature Medicine*, **25**, 1337-1340. <https://doi.org/10.1038/s41591-019-0548-6>
- [18] Niiranen, T.J., McCabe, E.L., Larson, M.G., Henglin, M., Lakdawala, N.K., Vasan, R.S., *et al.* (2017) Risk for Hypertension Crosses Generations in the Community: A Multi-Generational Cohort Study. *European Heart Journal*, **38**, 2300-2308. <https://doi.org/10.1093/eurheartj/ehx134>
- [19] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. <https://doi.org/10.1145/2939672.2939785>