心脏病风险的预测与影响因素的识别

盛焕新，程一鸣，陈贝宁

2024 年 1 月 12 日

摘要

本研究基于 Kaggle 心脏病风险数据，采用多种机器学习方法，综合应用深度学习模型，对心脏病预测进行了探讨。平衡化样本后，准确率提高至约 70%。复杂模型在预测准确度和 AUC 值上表现更佳。考虑心脏病的特点，短期生活习惯变化可能导致风险上升。在深度学习模型方面，我们提出了三个神经网络模型，准确度尚有提升空间。倾向得分回归揭示“是否肥胖”和“是否患糖尿病”对心脏病风险有显著影响。而从 Logistic 模型分析得出，胆固醇、糖尿病和睡眠时间在半球、大洲和国家层面呈现趋势性显著性。地理、文化和国家因素对心脏病风险产生显著差异，特定国家（如印度、意大利、日本）的生活方式和文化显著影响心血管健康。本研究为心脏病预测提供多角度洞察，尽管准确度仍有提升空间，但为理解心血管健康影响因素提供深刻见解，并根据模型结论提出了一些建议。

关键词： 机器学习; 数据挖掘; 心脏病预测; 因果探索.

目录

1 选题背景与数据情况 1

1.1 数据介绍 1

1.2 数据预处理 3

1.3 探索性数据分析 3

2 心脏病风险的机器学习模型预测 12

2.1 训练前数据准备 12

2.2 初步尝试 14

2.2.1 朴素贝叶斯和逻辑回归 14

2.2.2 基本树模型 15

2.2.3 Boost 类模型 18

2.3 模型调整 20

2.3.1 SMOTE 方法 20

2.3.2 朴素贝叶斯和逻辑回归 22

2.3.3 基本树模型 23

2.3.4 Boost 类算法 24

2.4 机器学习模型总结 25

2.5 深度学习模型 25

3 心脏病风险的影响因素识别 28

3.1 倾向得分回归 28

3.2 Logistic 模型 29

4 结论与建议 33

4.1 模型结论 33

4.2 对心脏病风险预防的建议 34

1 选题背景与数据情况

根据世界卫生组织 2023 年的全球卫生估算报告，心脏问题仍然是 21 世纪人类健康的头号杀手：2021 年有 2050 万人死于心血管疾病，这一数字约占全球死亡总人数的三分之一。而在 2050 万死于心血管疾病的人中，约有 900 万死于心脏病。但是，心脏病的发生一般较为突然，根据个人身体情况，提前评估心脏病风险并采取适当的防治措施对改善人们身体健康情况有比较重要的意义。

我们希望通过分析人们的基本信息、生活习惯、体检结果指标等记录，结合统计模型来识别心脏病高风险人群，从而提醒人们关注心脏健康，将心脏病治疗的关口前移，改善心脏病筛查的质量。

1.1 数据介绍

我们使用的数据来源于 Kaggle 竞赛网站 https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset/data。

我们使用的数据集包括 8763 条记录和 26 个变量。其中目标变量是对象是否被认为有心脏病风险，1 为有风险，0 为无风险。自变量的范围比较广泛，包括影响心脏健康的医学因素、受访对象的日常生活习惯以及受访对象的地理和经济情况等信息。具体的数据字典参见下表。

表 1: 数据集变量介绍

变量名	数据类型	取值范围
Heart Attack Risk（心脏病风险）	定性	0: 无风险; 1: 有风险
Patient ID（患者识别号）	/	/
Age（年龄）	定量	[18,90]
Sex（性别）	定性	0: 女性，1: 男性
Cholesterol（胆固醇）	定量	[120,400]

续下页

表 1: 续表

变量名	数据类型	取值范围
Blood Pressure（血压）	定量	舒张压 [60,110], 收缩压 [90,180]
Heart Rate（心率）	定量	[40,110]
Diabetes（是否患糖尿病）	定性	0: 无糖尿病; 1: 有糖尿病
Family History（是否有心脏病家族史）	定性	0: 无; 1: 有
Smoking（是否抽烟）	定性	0: 无; 1: 有
Obesity（是否肥胖）	定性	0: 无; 1: 有
Alcohol Consumption（饮酒程度）	定性	无/轻微/中等/重
Exercise Hours Per Week（周运动时间）	定量	[0,20]
Diet（饮食习惯）	定性	健康/平均/不健康
Previous Heart Problems（心脏病病史）	定性	0: 无; 1: 有
Medication Use（是否用药）	定性	0: 无; 1: 有
Stress Level（压力水平）	定量	[1,10]
Sedentary Hours Per Day（每日坐姿时长）	定量	[0,24]
Income（收入水平）	定量	[0,300000]
BMI（身体质量指数）	定量	[18,40]
Triglycerides（甘油三酯）	定量	[30,800]
Physical Activity Days Per Week（每周运动天数）	定性	0,1,2…7
Sleep Hours Per Day（每日睡眠时长）	定量	[4,10]
Country（患者所在国家）	定性	中国等 20 个国家

续下页

表 1: 续表

变量名	数据类型	取值范围
Continent（患者所在大洲）	定性	除南极洲外所有洲
Hemisphere（患者所在半球）	定性	南半球；北半球

1.2 数据预处理

为了更好地获取数据中的信息，我们需要对数据进行预处理。主要分为以下几步。

首先，检查数据无缺失值，并筛选除去与模型训练无关的患者 ID 等变量。

为了提升数据的稳定性，降低噪声影响，我们对年龄、周运动时间和收入等变量进行分箱处理。我们按照皮尤研究中心的方法，将受访对象根据出生的时代分为五组，对应到年龄，分别是 78-95 岁，59-77 岁，43-58 岁，27-42 岁，11-26 岁和 10 岁及以下。按照美国心脏协会（American Heart Association）推荐的运动水平，我们按照周运动时间将受访对象分为低运动强度组（每周运动 0-119 分钟），正常运动强度组（每周运动 120-180 分钟）和高运动强度组（每周运动时间超过 180 分钟）。我们还将受访对象分为低收入组（收入低于 30000），中等收入组（收入在 30000-120000 之间）和高收入组（收入高于 120000）。

此外，由于血压数据以字符串形式记录，我们根据衡量心脏健康的一般做法，将记录的血压数据分成收缩压和舒张压，并最终利用收缩压和舒张压的差值（又称脉压）作为衡量受访对象血压水平的变量。

1.3 探索性数据分析

首先我们了解一下样本标签和各影响因素的分布情况。

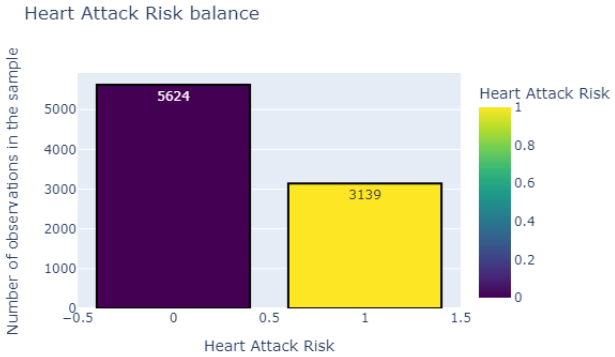


图 1: 样本的标签分布

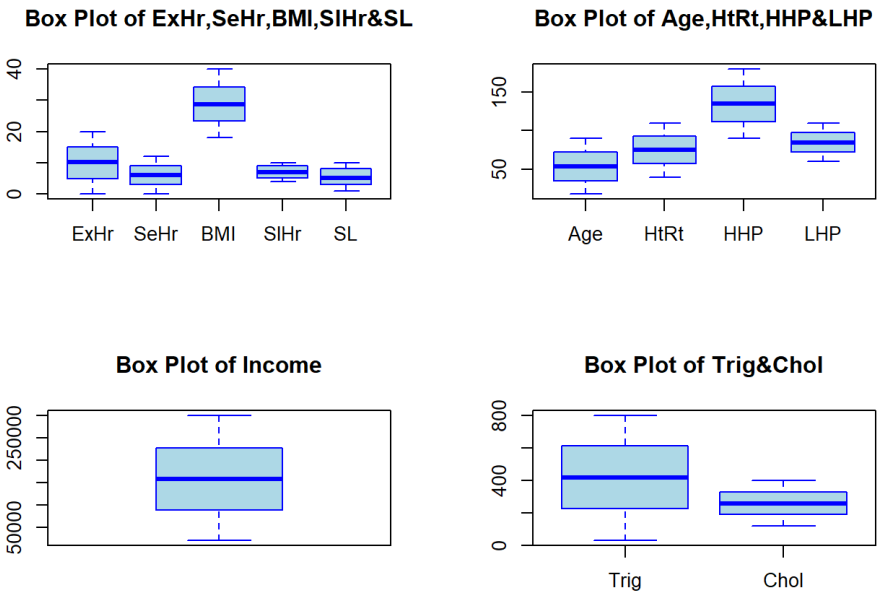


图 2: 数值型变量的分布

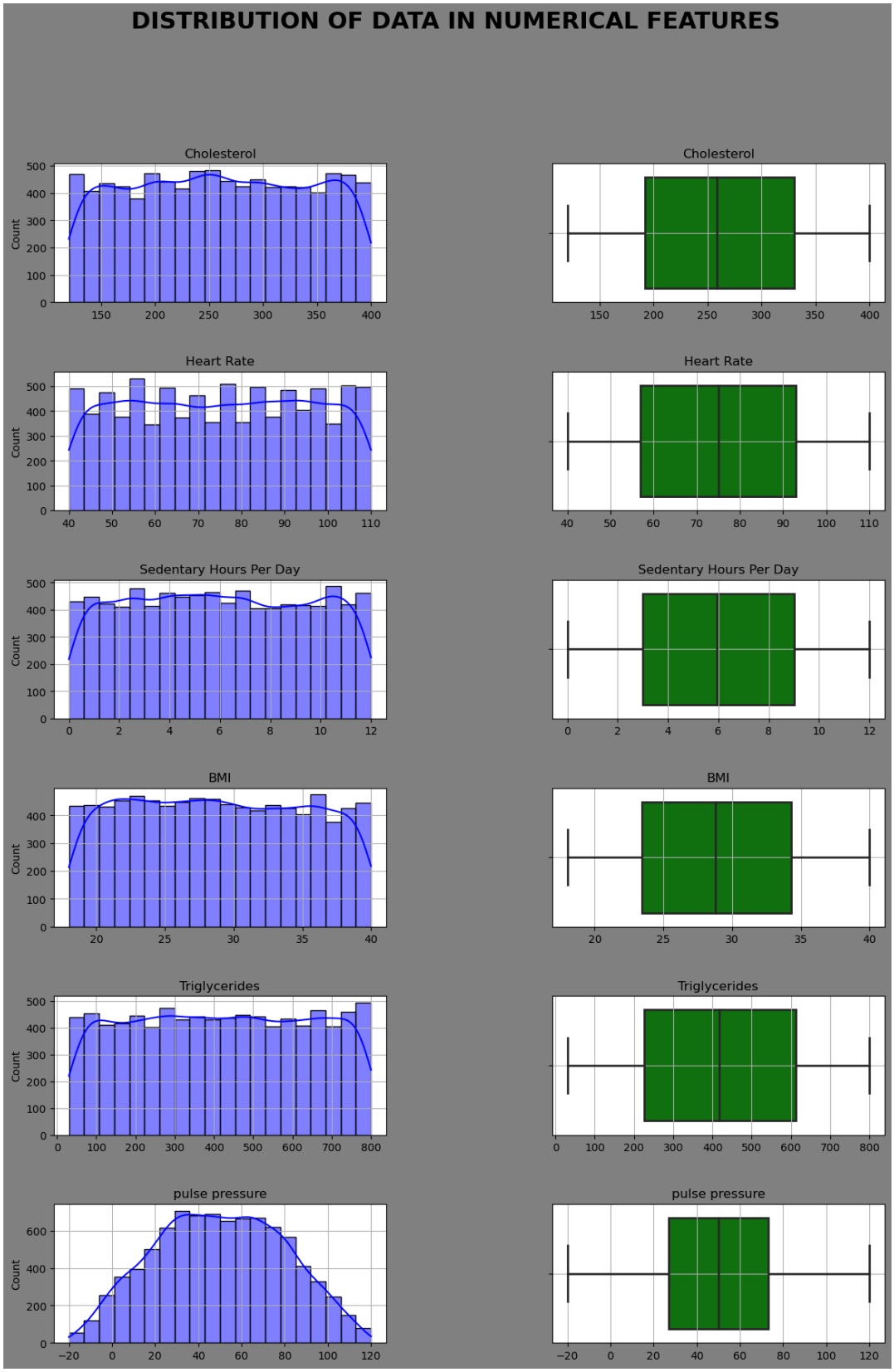

DISTRIBUTION OF DATA IN NUMERICAL FEATURES
Cholesterol
Heart Rate
Sedentary Hours Per Day
BMI
Triglycerides
pulse pressure
Count

图 3: 数值型变量的分布

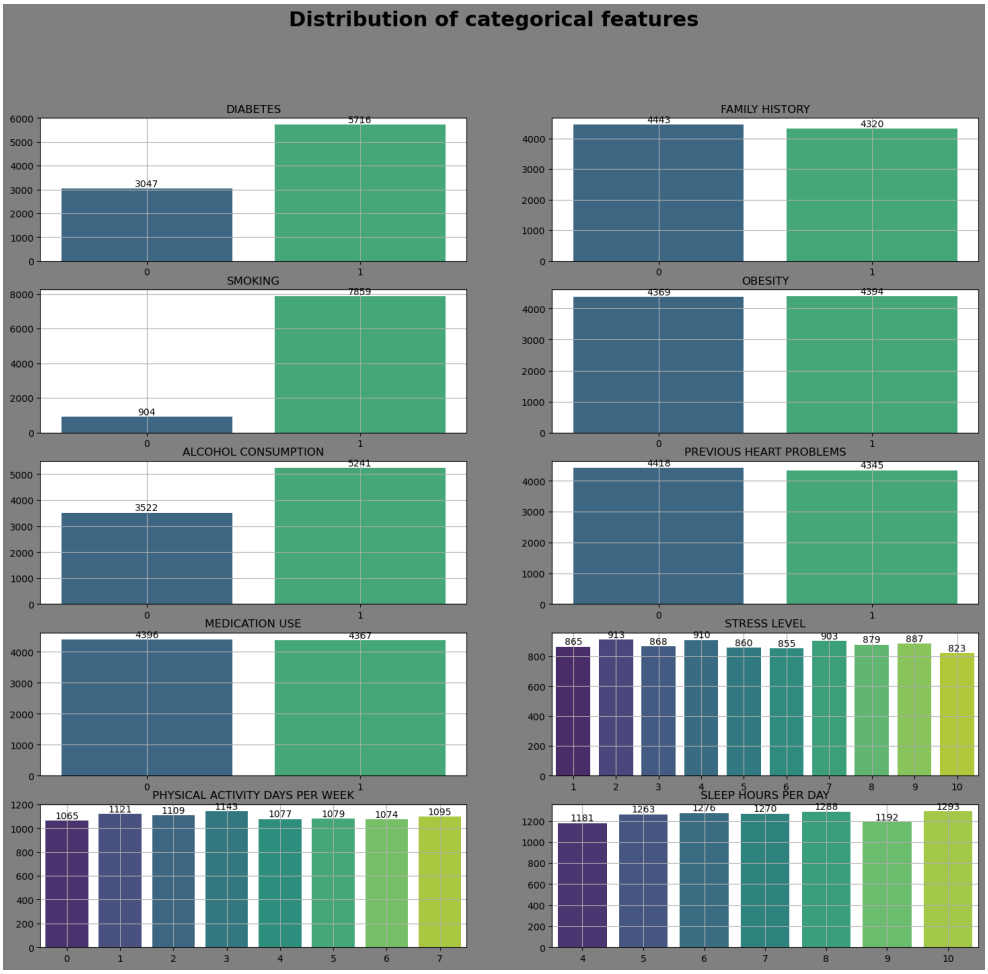


图 4: 类别型变量的分布

通过分析自变量和因变量的分布，我们发现：

1. 样本中标记为无心脏病风险的样本占总体的约 64%，样本存在一定程度的不平衡现象。

2. 数值型变量计算的尺度不一致，构建模型时直接使用可能会带来不稳定因素。

3. 数值型和类别型变量的分布都较为平均，没有明显的差异，提示模型构建时应考虑混杂因素的干扰。

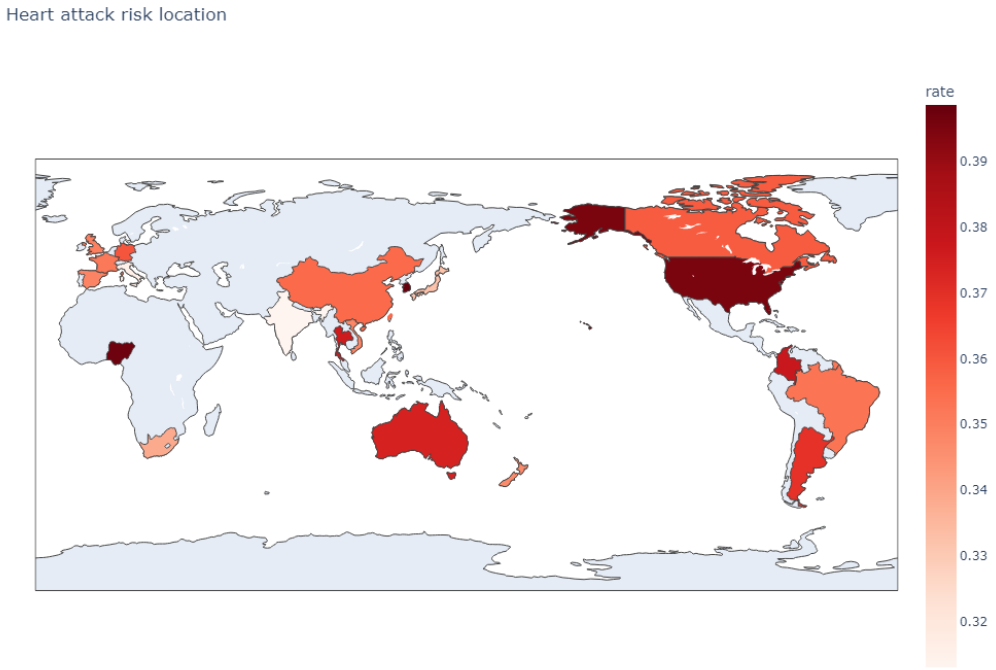


图 5: 心脏病风险的地域分布

通过分析不同地区心脏病风险的识别率，发现美洲和欧洲、亚洲等地风险概率存在差别，可以考虑进一步识别地区的环境因素对心脏病风险的影响。

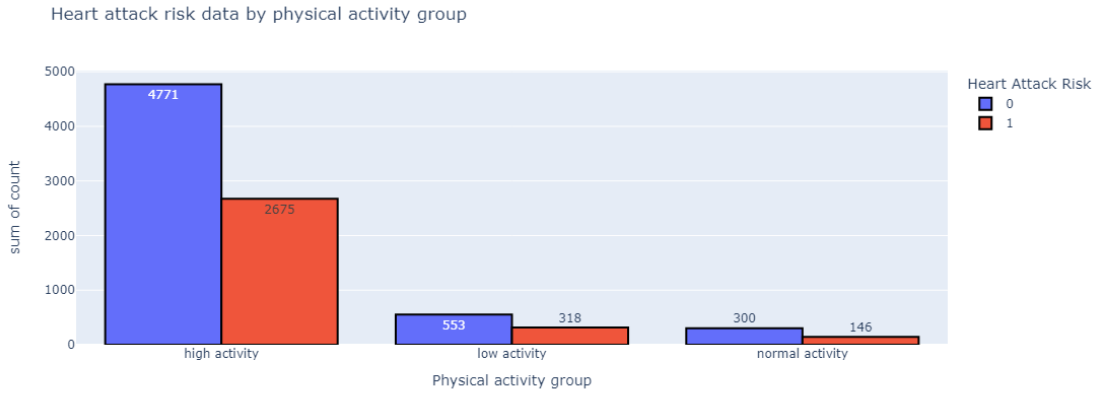


图 6: 按运动习惯分组的心脏病风险分布

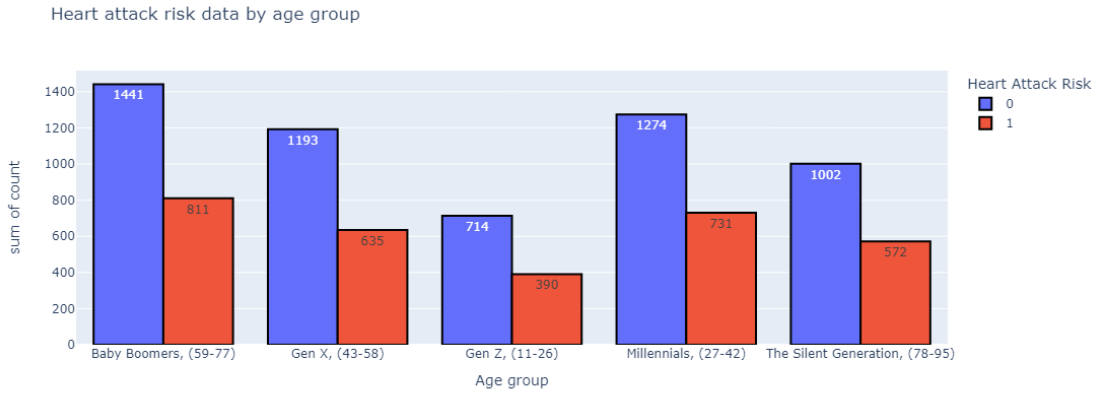


图 7: 按年龄分组的心脏病风险分布

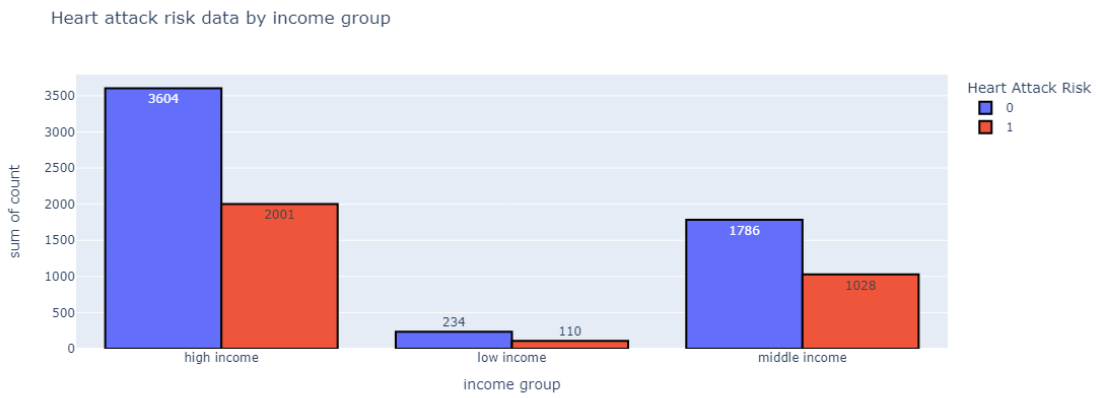


图 8: 按收入水平分组的心脏病风险分布

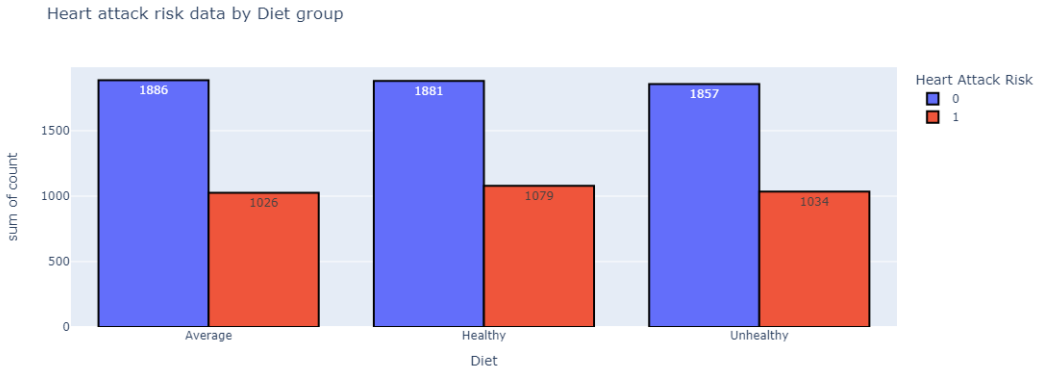


图 9: 按饮食习惯分组的心脏病风险分布

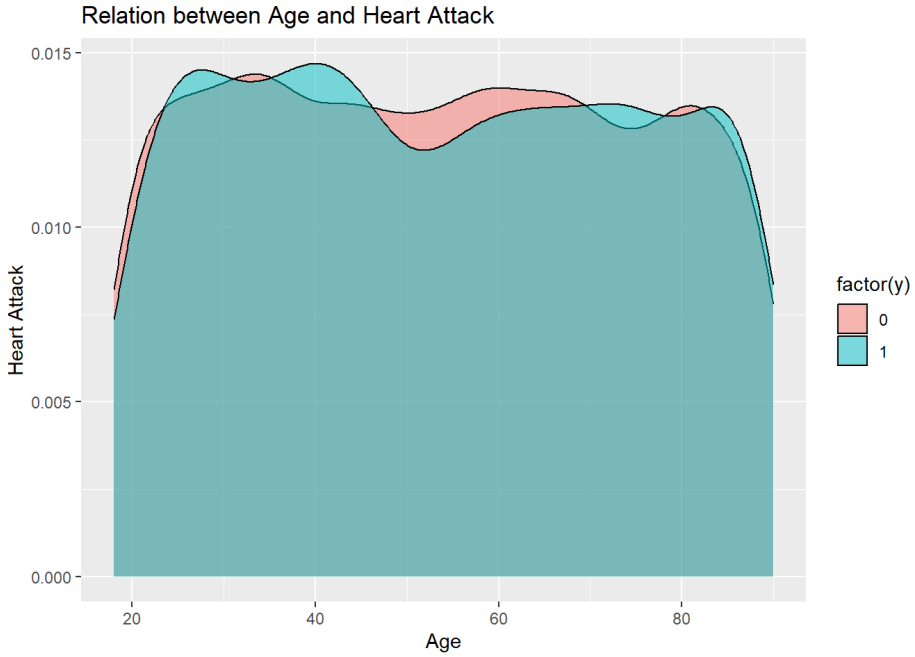


图 10: 不同心脏病风险的年龄分布

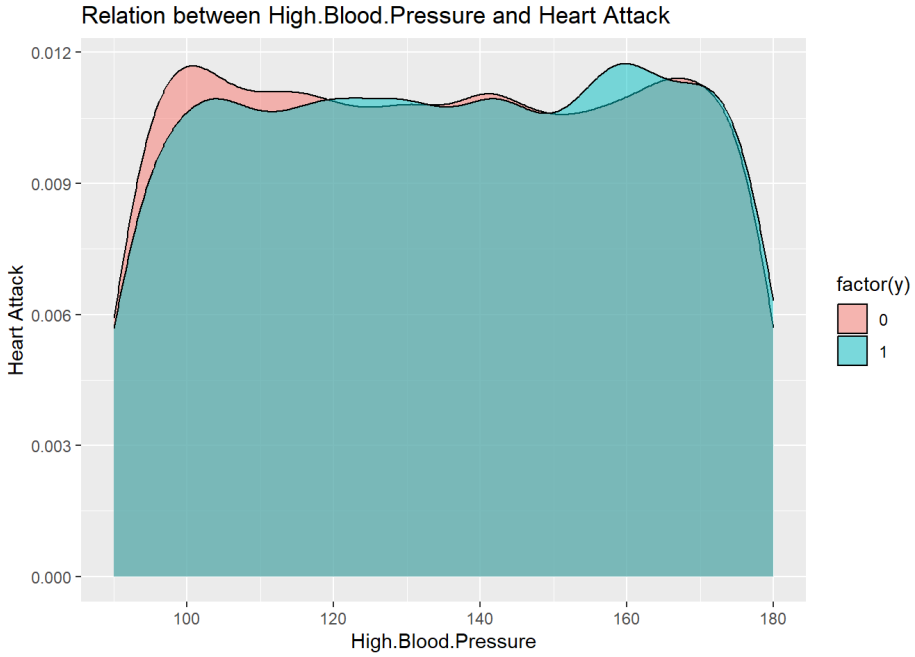


图 11: 不同心脏病风险的收缩压分布

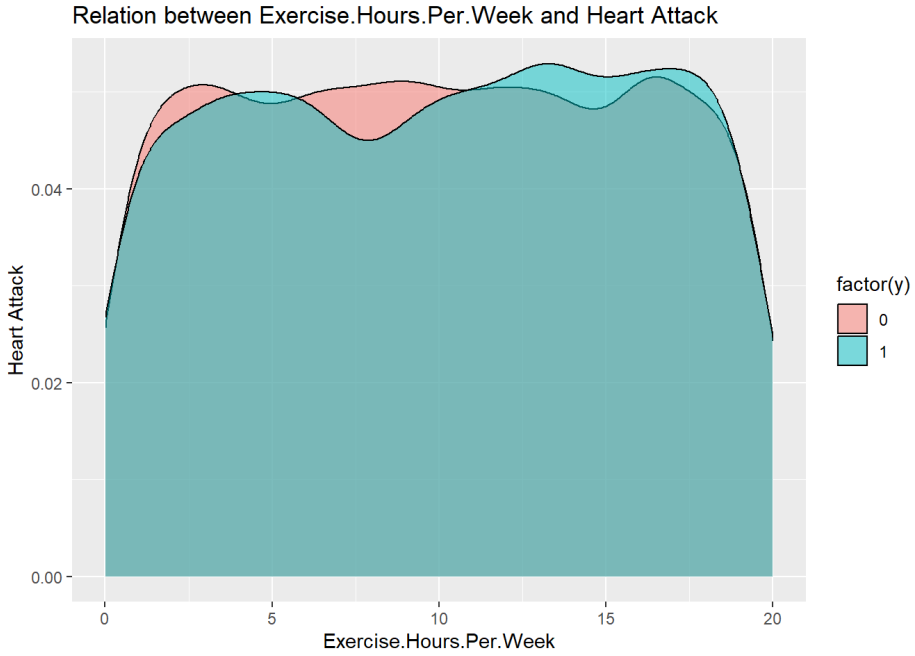


图 12: 不同心脏病风险的运动时间分布

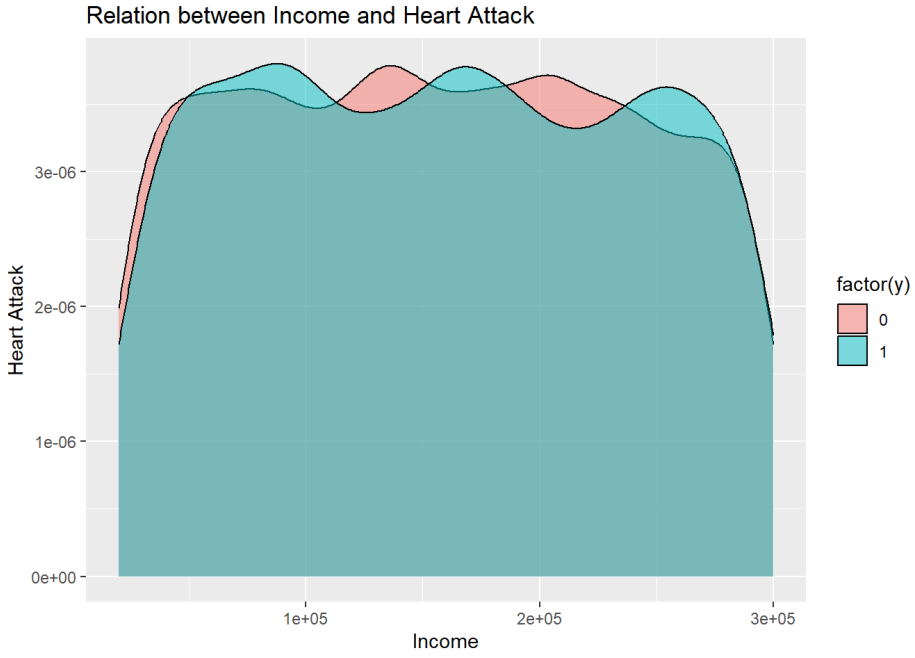


图 13: 不同心脏病风险的收入分布

通过分析心脏病风险和不同类别型变量的关系，我们发现各组中心脏病风险概率和总体基本一致，不能直接从某些变量的比较中直接得到影响因素，需要进一步识别影响因素。通过分析不同心脏病风险组别中连续型变量的分布，可以看到两组间分布并没有显著差别，说明数据高度线性不可分。这使得之后的模型训练过程更加复杂。

2 心脏病风险的机器学习模型预测

在这一部分，我们希望能够借助现有的机器学习模型并加以修改，利用现有数据，对心脏病风险进行预测，从数据角度帮助识别潜在的心脏病风险。

我们在这一步分主要使用了朴素贝叶斯、逻辑回归、决策树、随机森林、Gradient Boost 和 XGBoost 等算法

2.1 训练前数据准备

结合前述数据的特征，我们在训练模型前对数据做了如下处理来降低噪声和可能的过拟合的影响。

- 对数据进行独热编码，最终得到 59 个特征
- 检查数据的相关性
- 使用 Pearson 和 ANOVA 检验筛选出较为重要的 25 个变量
- 使用 MinMaxScaler 对数据进行标准化
- 按照 8:2 的比例将数据划分为训练集和测试集

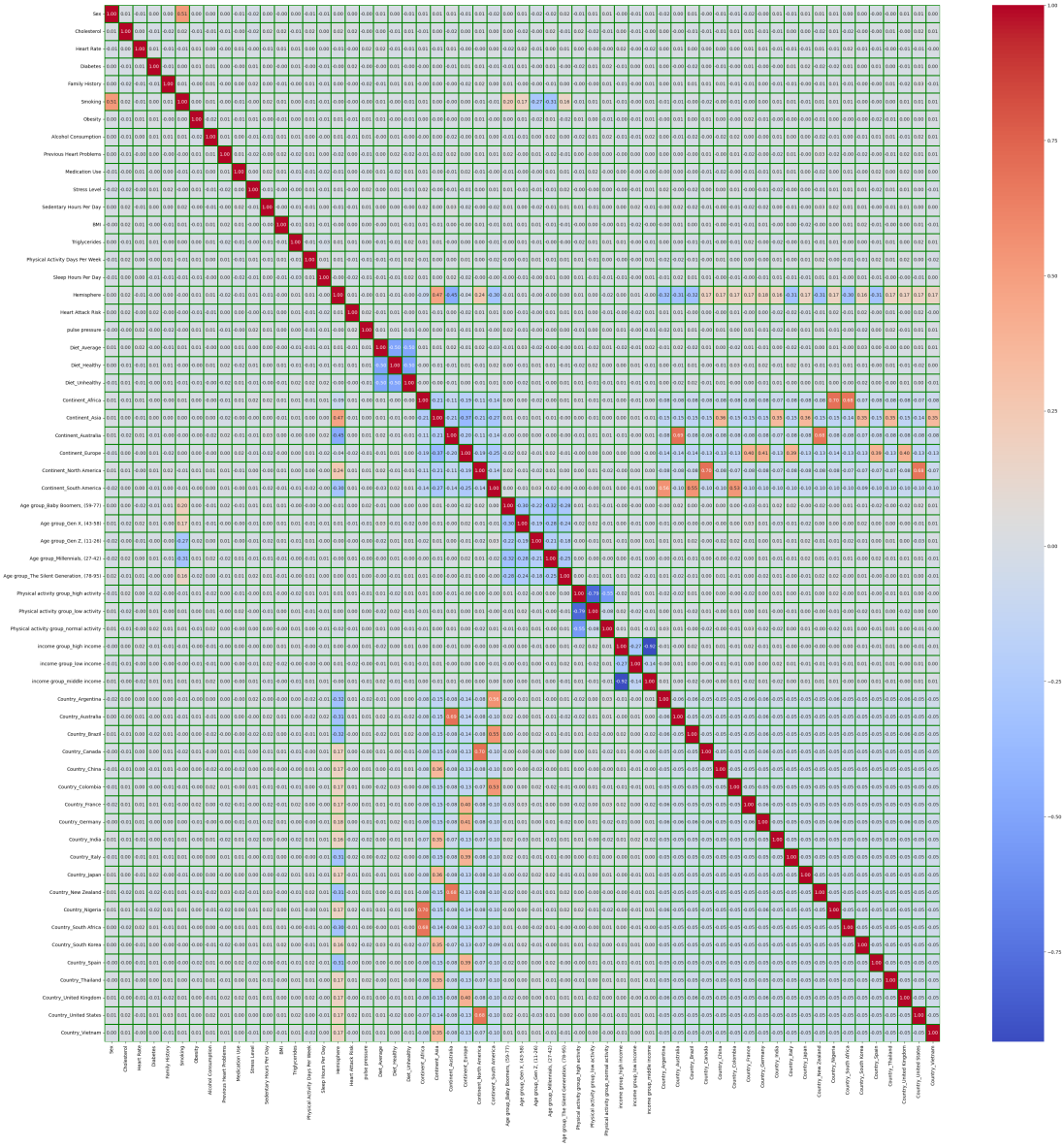


图 14: 数据相关性热力图

分析相关性热力图发现数据之间相关性绝大部分都接近于 0，显著非零部分为变量设置时重叠的部分。

2.2 初步尝试

我们在这一步分主要使用了朴素贝叶斯、逻辑回归、决策树、随机森林、Gradient Boost 和 XGBoost 等算法

2.2.1 朴素贝叶斯和逻辑回归

使用较为简单的朴素贝叶斯和逻辑回归模型，分析预测结果阵可以看到，总体的准确率约 64%，与全样本中负样本比例接近，混淆矩阵也提示这些方法将所有样本均预测为负样本，与识别心脏病风险的目标不符。

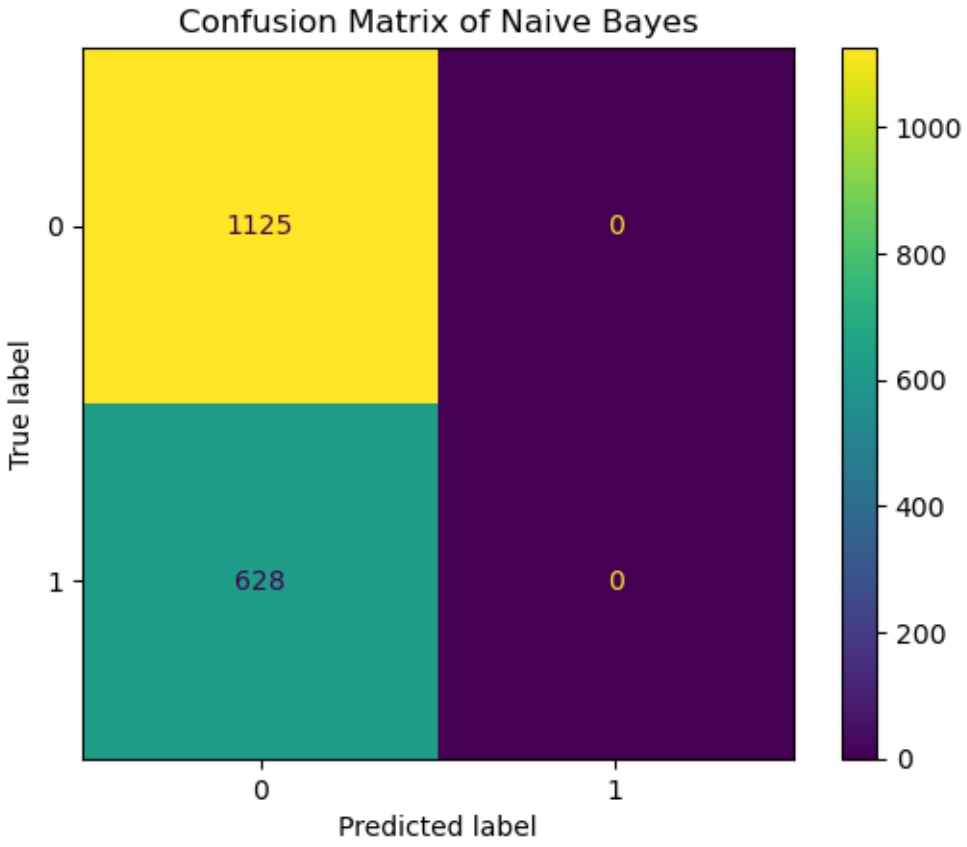


图 15: 朴素贝叶斯混淆矩阵

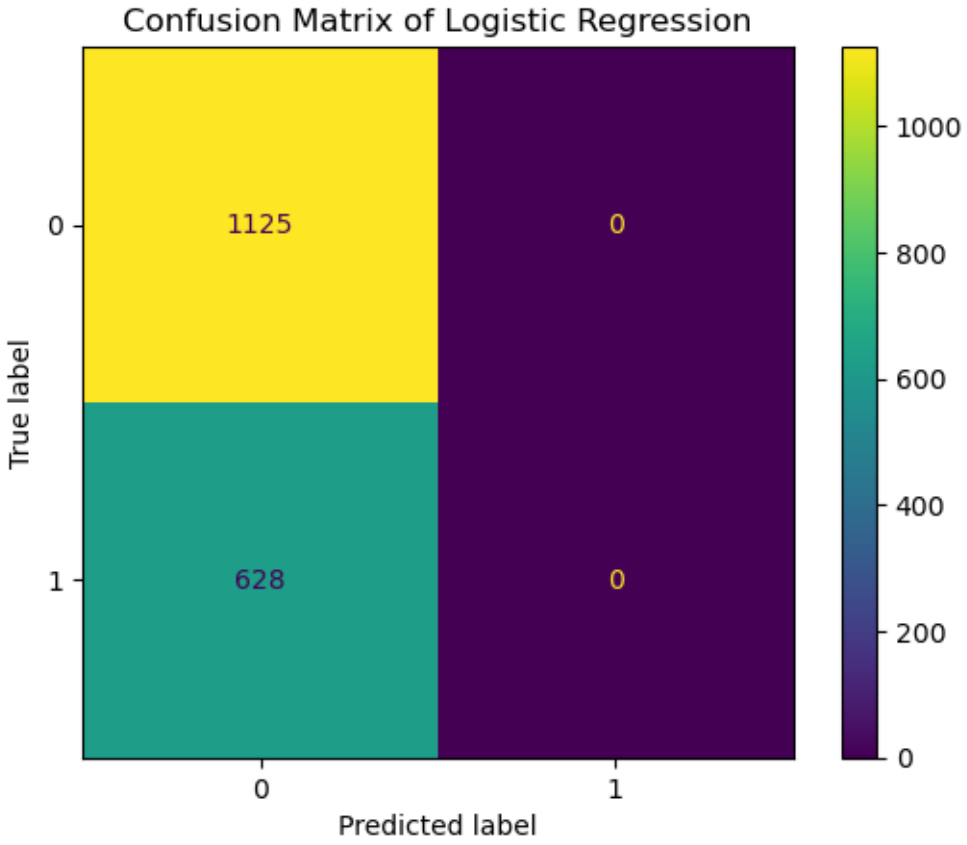


图 16: 逻辑回归混淆矩阵

2.2.2 基本树模型

使用决策树和随机森林等树结构的基本模型进行预测,准确率仍为接近 64%,但能够识别出一些正样本。

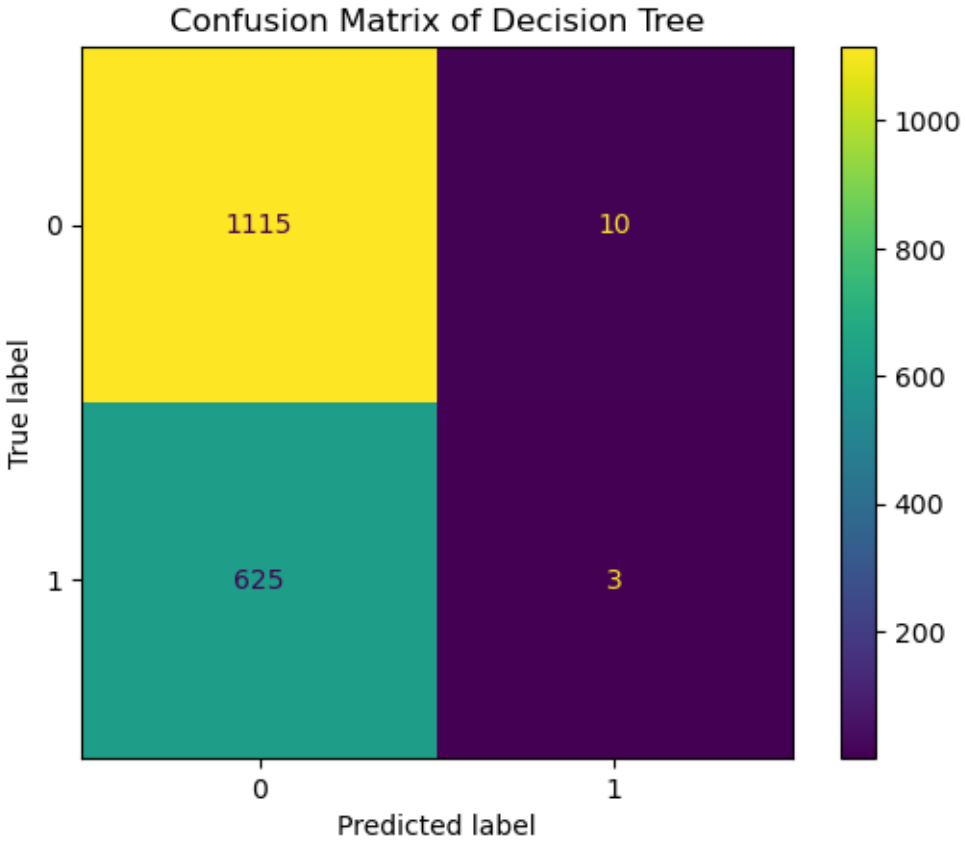


图 17: 决策树模型混淆矩阵

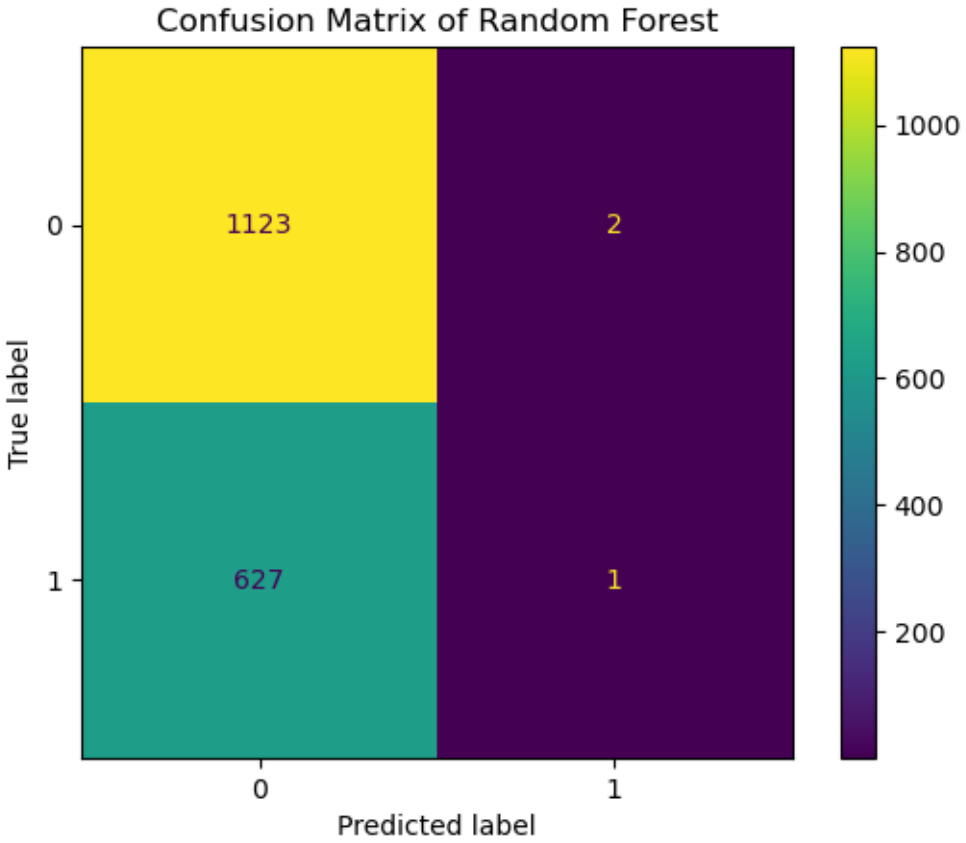


图 18: 随机森林模型混淆矩阵

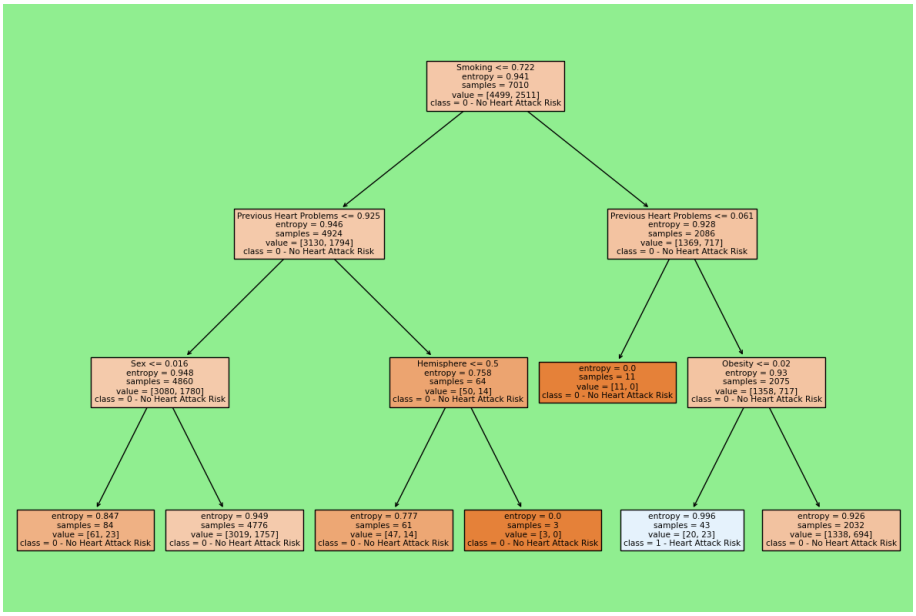


图 19: 根据参数绘制的决策树样例

2.2.3 Boost 类模型

使用基于树模型改进得到的 Boost 类模型中 Gradient Boost 和 XGBoost 模型进行训练，混淆矩阵提示识别到了更多的正样本，但是总体的准确率改进不大，依然在 64% 左右。

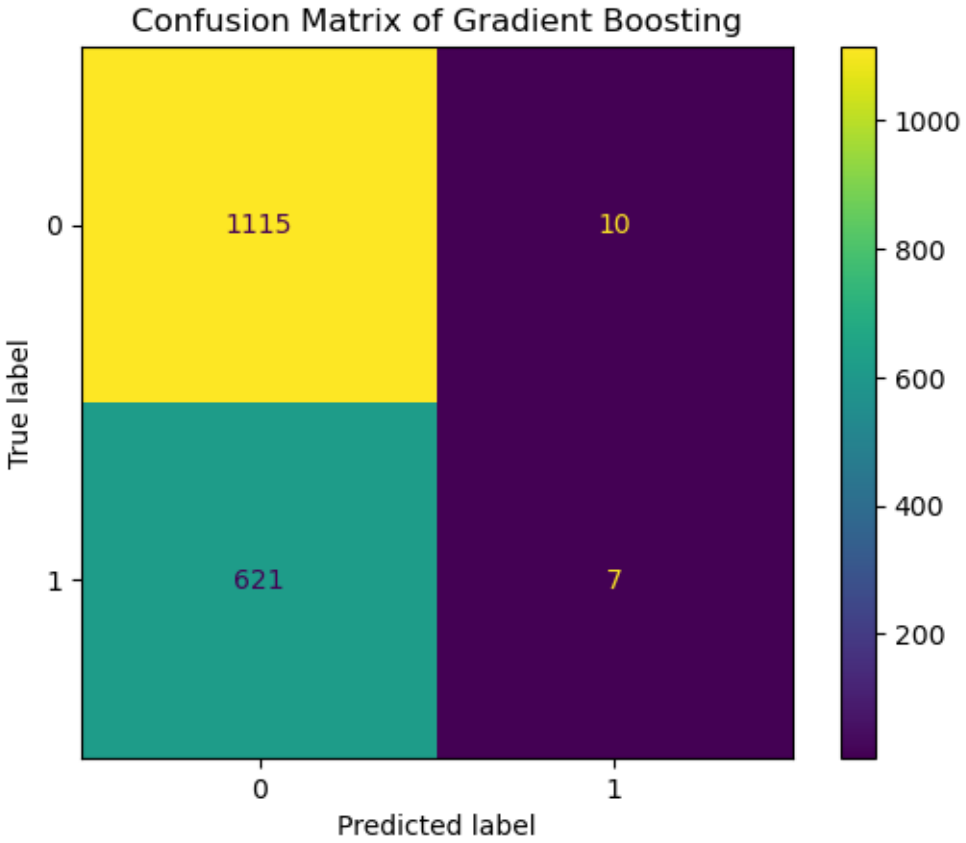


图 20: Gradient Boost 模型混淆矩阵

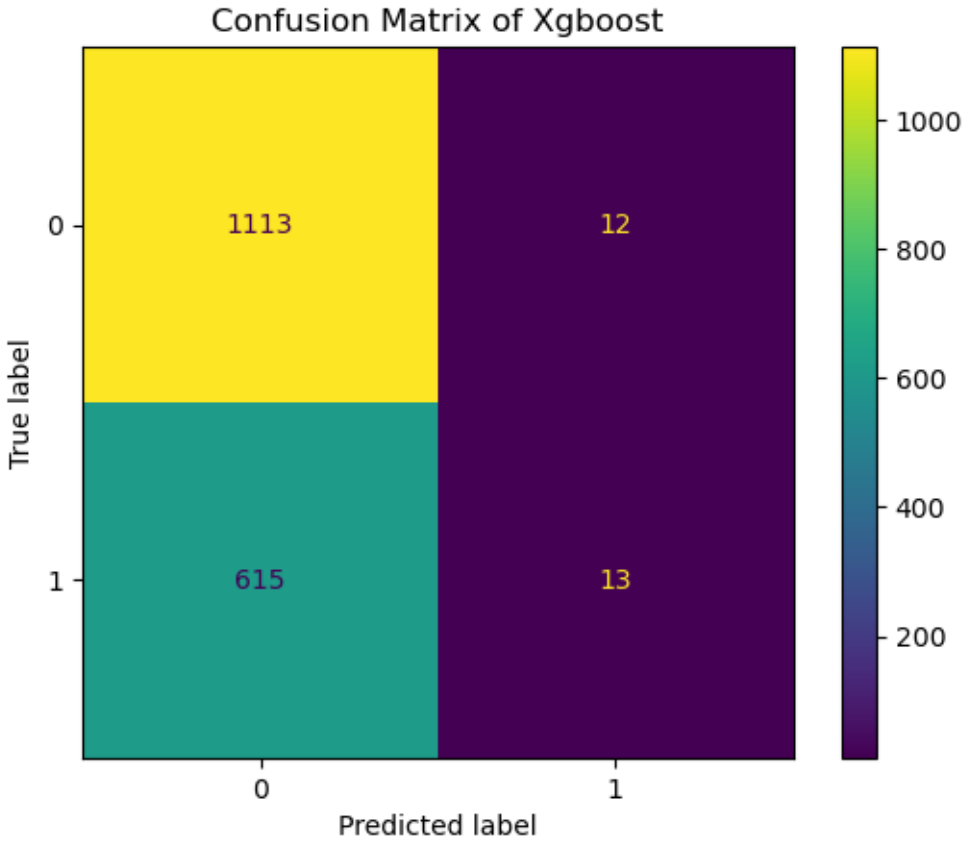


图 21: XGBoost 模型混淆矩阵

2.3 模型调整

在调整的数据集上直接应用机器学习模型表现并不好，而我们注意到样本因变量存在一定程度的不平衡特征，因此我们希望构建一个正负样本平衡的新数据集，利用新数据集构建模型并检验。在之后的模型构建中，我们使用了网络搜索和 optuna 方法来调整参数，并且采用 AUC 值来评价模型的预测效果。

2.3.1 SMOTE 方法

在这里我们使用 SMOTE 采样来平衡正负样本。

SMOTE(Synthesized Minority Oversampling Technique) 方法是以每个样本

点的 k 个最近邻样本点为依据，随机选择其中 N 个邻近点，建立邻近点到样本点的向量，在向量上利用 [0,1] 间的随机因子选取新的数据点作为样本点，从而达到合成数据的目的。简要的说就是合成新的少数类样本来平衡数据集。

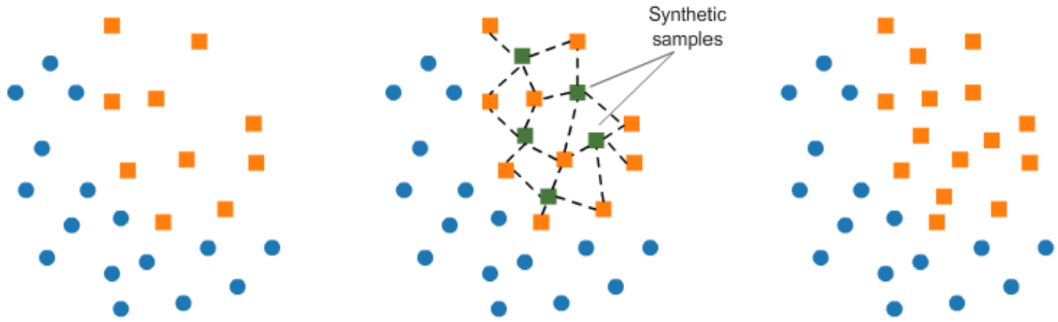


图 22: SMOTE 方法示意

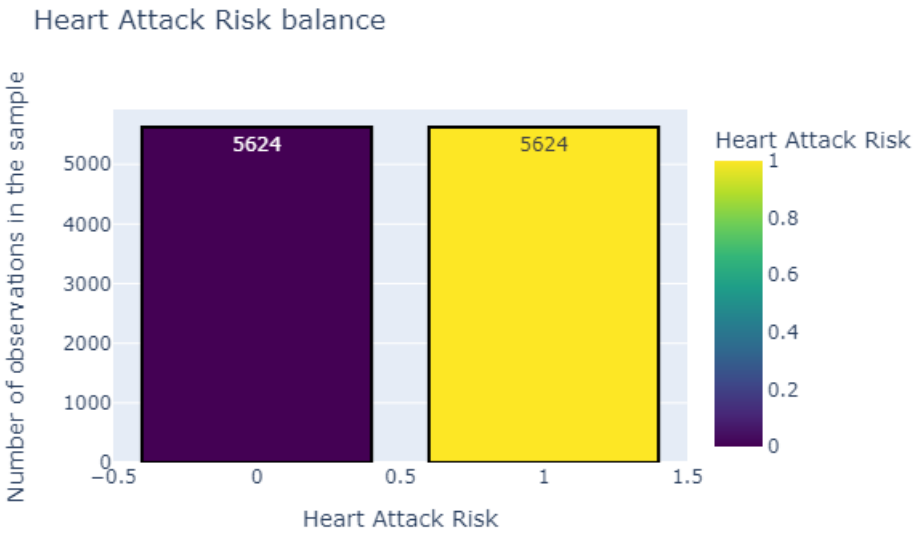
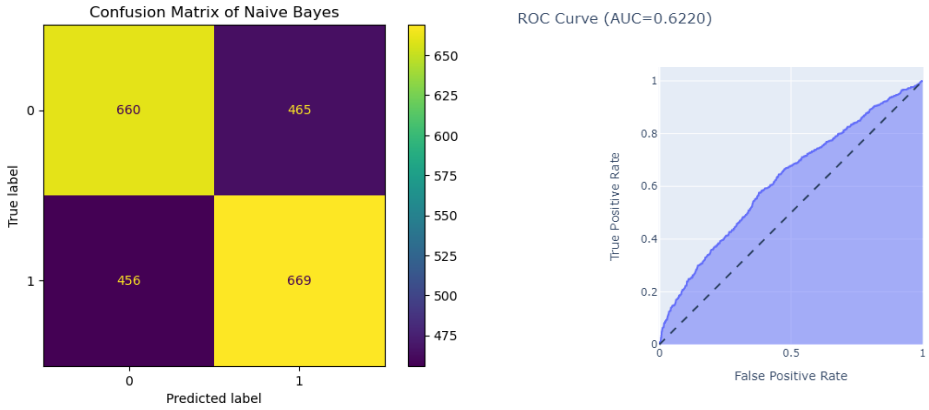


图 23: 平衡后的样本

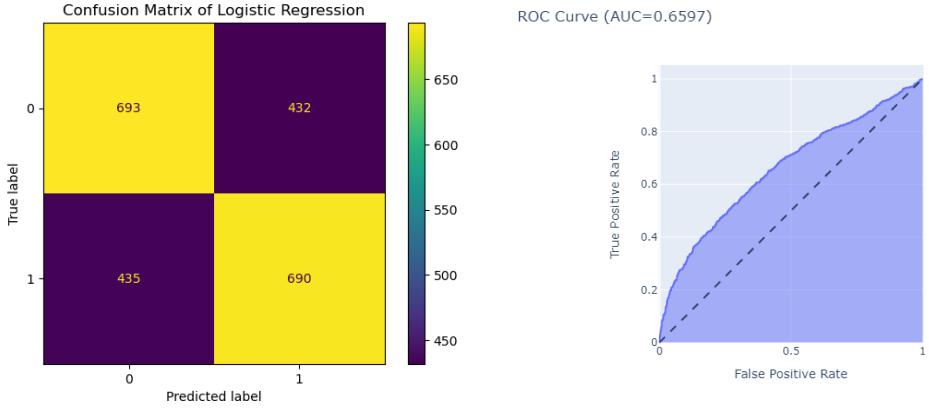
2.3.2 朴素贝叶斯和逻辑回归

平衡样本后，模型评价结果显示正样本的预测准确率上升，负样本的预测准确率有所下降，总体准确率仍约 60%。不过 AUC 值相比未经样本调整的模型有了提升，由接近 0.5 提升到约 0.62。



(a) 混淆矩阵 (b) ROC 曲线

图 24: 改进后朴素贝叶斯模型评价

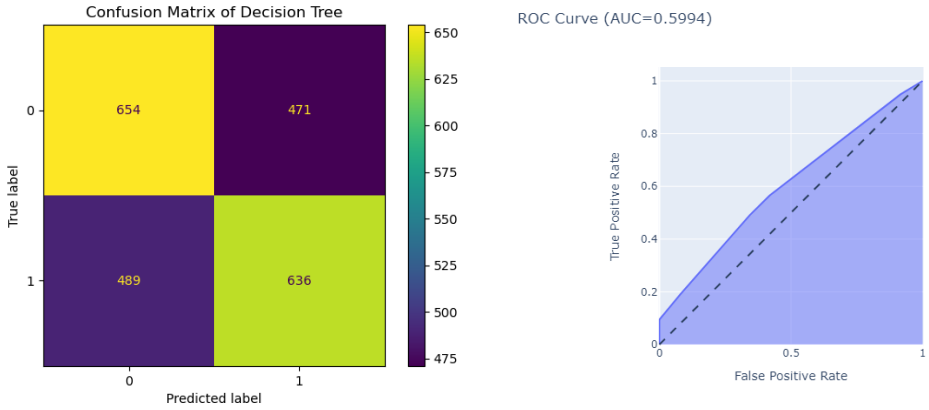


(a) 混淆矩阵 (b) ROC 曲线

图 25: 改进后逻辑回归模型评价

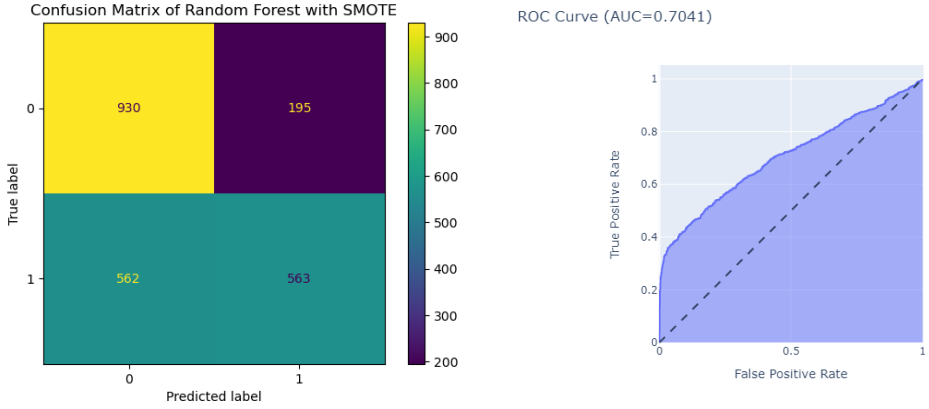
2.3.3 基本树模型

经过样本调整后，决策树模型的 AUC 值也有所提升。而且随机森林模型的预测精度超过 64%，达到 66%，AUC 值也显著提升。说明样本调整对预测效果具有改善作用。



(a) 混淆矩阵 (b) ROC 曲线

图 26: 改进后决策树模型评价

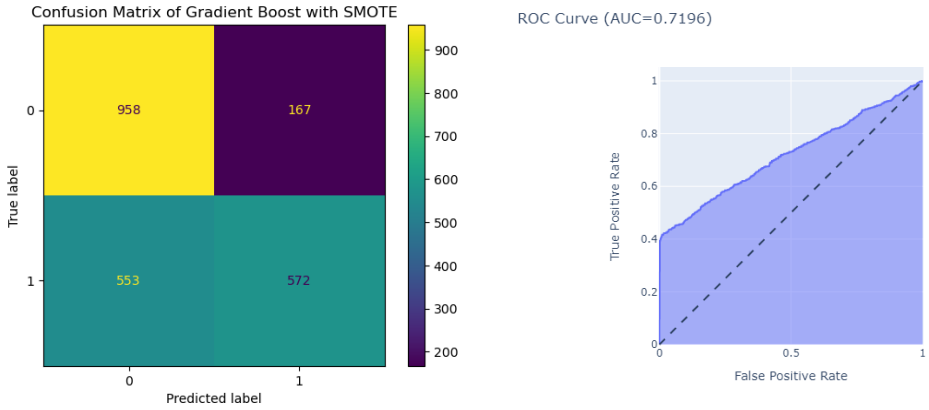


(a) 混淆矩阵 (b) ROC 曲线

图 27: 改进后随机森林模型评价

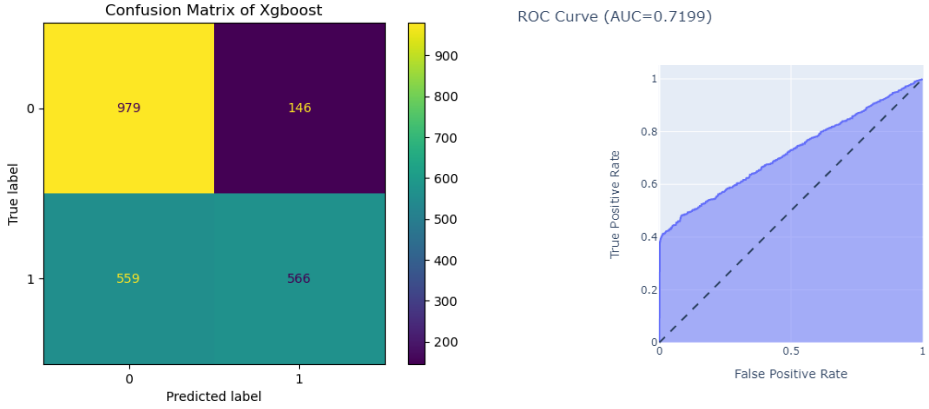
2.3.4 Boost 类算法

Boost 类算法表现相对更好。与前述模型相比，不仅保持了一定的负样本预测准确率，也改善了正样本的预测准确率，总体预测准确率达到 70%。而且 AUC 值也有较好表现，达到约 0.72。



(a) 混淆矩阵 (b) ROC 曲线

图 28: 改进后 Gradient Boost 模型评价



(a) 混淆矩阵 (b) ROC 曲线

图 29: 改进后 XGBoost 模型评价

2.4 机器学习模型总结

通过比较不同方法的机器学习模型，我们发现总体准确度最好也只能达到约64%，在考虑平衡化样本后，总体准确率达到约 70%。同时，使用复杂模型比简单模型在预测准确度和 AUC 值的表现上均有所提升。针对预测准确率相对不高的问题，我们考虑这与心脏病的特点有关。心脏病发病具有一定的随机性与突然性，良好的作息习惯和体检指标仅能在概率上降低心脏病风险，但不能排除偶然因素的影响，有时短时间内生活习惯上的剧烈变化也会直接导致心脏病风险上升。因此，我们希望进一步识别造成心脏病风险的具体因素，帮助患者针对性规避可能的心脏病风险。不过，在因素识别之前，我们还尝试了深度学习方法进行预测准确度的改进。

2.5 深度学习模型

本任务是一个典型的二分类问题，旨在通过给定的输入数据来预测目标的类别。我们将数据集划分为 8:2 的训练集和测试集。在模型训练之前，我们对输入数据进行了标准化处理，使用了 StandardScaler 来确保所有特征具有相似的尺度。这有助于优化梯度下降算法的性能，提高模型训练的稳定性。

我们设计了一个深度神经网络模型，称之为 SHXNN。该模型包含 10 个全连接层，每一层都包括批量归一化、激活函数（ReLU）和丢弃层（Dropout），这些层的引入旨在增强模型的泛化能力和防止过拟合。模型的最后一层使用 Sigmoid 激活函数，输出一个 0 到 1 之间的值，代表样本属于正类别的概率。

我们选择了以下参数设置用于构建 SHXNN 模型：

- 输入层维度：`input_size = X_train_scaled.shape[1]`
- 隐藏层维度：128, 256, 512, 1024, 512, 256, 128, 64, 16
- Dropout 率：0.1
- 激活函数：ReLU 在每个隐藏层

- 输出层：Sigmoid 激活函数
- 权重初始化：使用了现代的权重初始化方法，如 Kaiming 和 Xavier 初始化

模型的训练采用了二元交叉熵损失函数（BCEWithLogitsLoss）和 Adam 优化器。为了更好地调整学习率，我们使用了指数衰减学习率调度器。这有助于在训练过程中逐渐减小学习率，提高模型在后期的收敛性。在模型评估阶段，我们选择了准确率（Accuracy）和召回率（Recall）作为评估指标。准确率表示模型正确分类的样本比例，而召回率衡量了模型正确预测正类别样本的能力，尤其关注于未被模型漏掉的正类别样本。

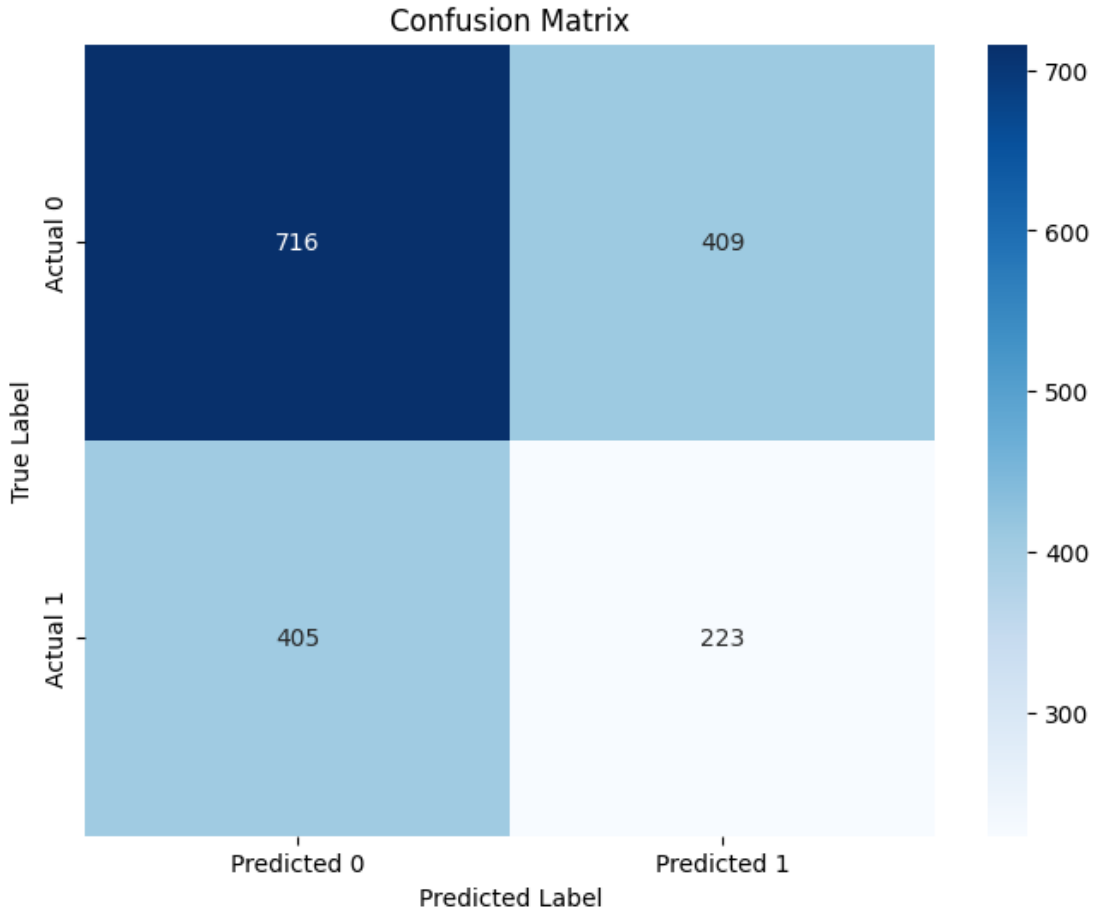


图 30: 不带 SMOTE 和 DROPOUT 的 10 层神经网络模型

为了处理不平衡的类别分布，我们使用了 SMOTE 技术对训练集进行过采样。SMOTE 通过在特征空间中生成合成样本，从而增加少数类样本的数量，使得两个类别的样本数相对平衡

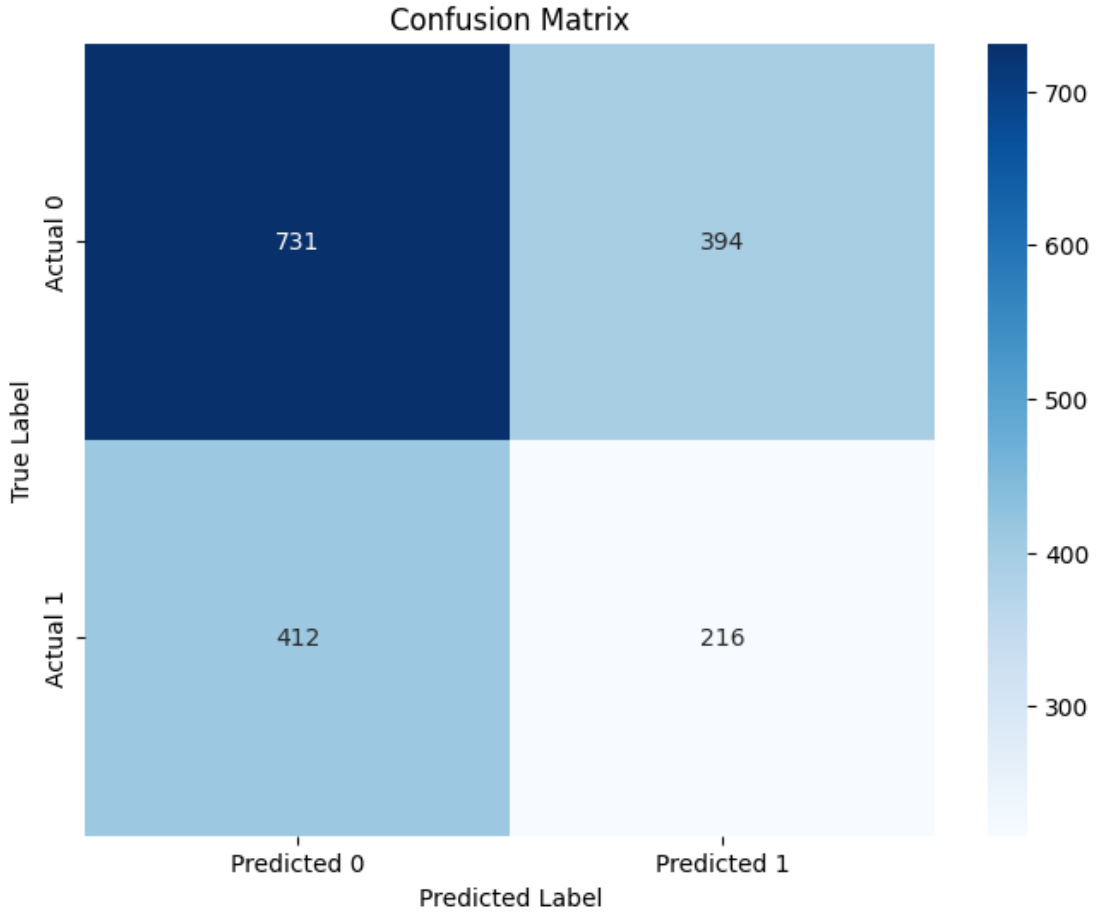


图 31: 带 SMOTE 不带 DROPOUT 的 10 层神经网络模型

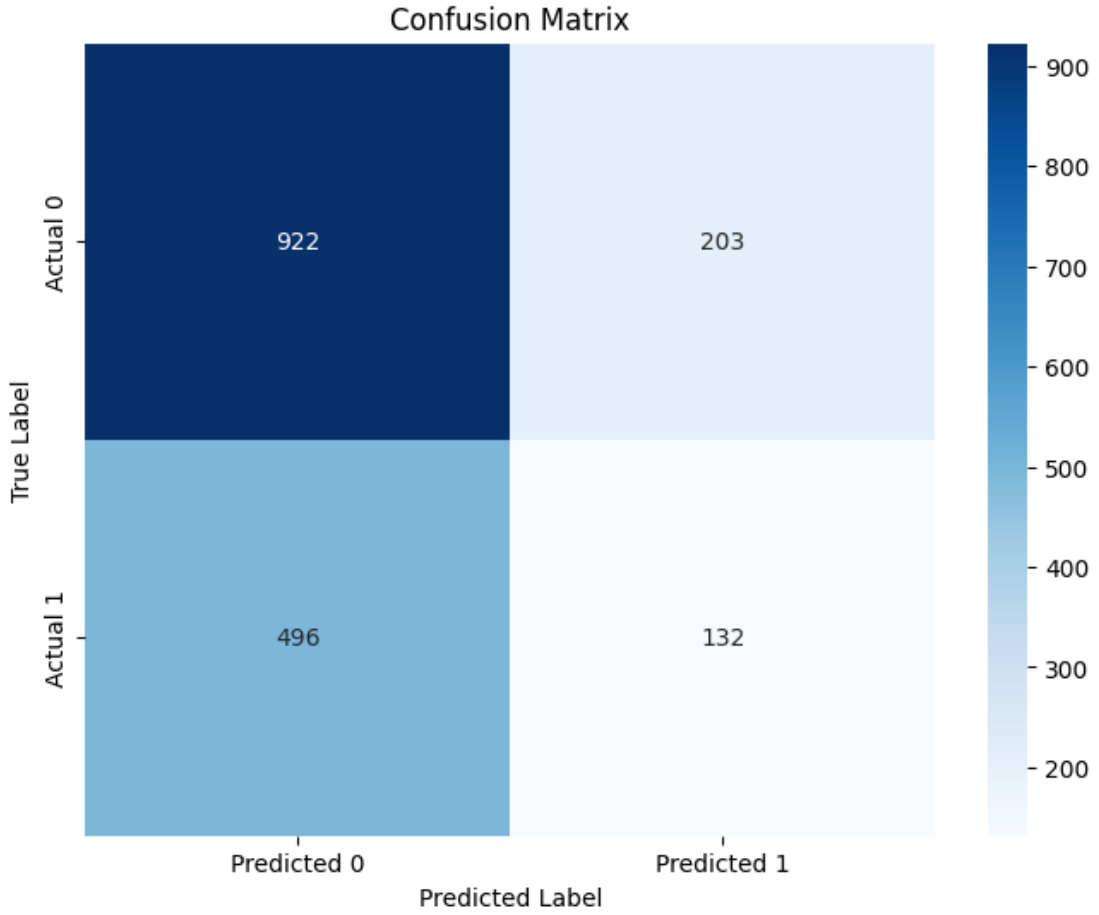


图 32: SMOTE+0.1DROPOUT 的 10 层神经网络模型

3 心脏病风险的影响因素识别

3.1 倾向得分回归

为了能够更精准地识别分类变量对心脏病风险预测的影响，需要在目标分类变量之外控制混杂因素一致。借鉴倾向得分的思想，我们将患者特征中每个分类变量作为处理变量，计算样本的倾向得分。其中，倾向得分是每个样本被选择为

实验组的概率的预测值，

$$T_i = \alpha + \sum_{k=1}^{K} X_k + \epsilon_i$$

使用逻辑回归进行计算，$\hat{T_i}$ 即为倾向得分预测值。得到倾向得分后，做回归

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 \hat{T_i} + \epsilon_i$$

其中 T_i 是样本是否为处理组，$\hat{T_i}$ 是倾向得分。得到回归结果后，对 $\hat{\beta}_1$ 进行检验，通过检验结果可以推断和识别处理变量的影响。

最终识别得到“是否肥胖”和“是否患糖尿病”两个分类变量对心脏病风险有显著影响。

3.2 Logistic 模型

本节将深入探讨三个 Logistic 回归模型的结果，分别针对 Hemisphere-wise、Continent-wise 和 Country-wise 层面。这些模型旨在揭示心血管健康的影响因素，并通过对大洲、国家和半球的分析，提供了多层次、全面的视角。通过对模型系数、p 值和整体显著性的分析，我们将深入了解各因素对心血管疾病风险的潜在影响。

首先，我们将对 Hemisphere-wise 模型进行详细讨论，探讨在半球层面上，年龄、性别、胆固醇、糖尿病、睡眠时长等变量对心血管健康的贡献。接着，Continent-wise 模型的分析将进一步拓展我们的视野，关注大洲层面的地理因素对心血管健康的可能影响。最后，我们将深入研究 Country-wise 模型，聚焦具体国家级别的因素，如印度、意大利、日本等，以揭示它们在心血管疾病发生中的独特角色。

通过这一系列的模型分析，我们旨在为制定定制化的心血管健康干预措施提供深刻理解，同时强调在解释模型结果时需综合考虑统计显著性、实际效果和领域知识。这有助于确保我们的报告不仅具有科学严谨性，而且能够为决策者和健康专业人员提供有实质性意义的洞见。

表 2: Logistics 结构模型：Hemisphere-wise

变量	估计值	标准误差	p 值
(Intercept)	-7.685e-01	2.815e-01	0.00633**
Age	1.174e-03	1.176e-03	0.31807
Sex	4.406e-02	5.842e-02	0.45073
Cholesterol	**5.010e-04**	**2.765e-04**	**0.07003.**
Heart.Rate	-4.569e-04	1.087e-03	0.67411
Diabetes	**7.827e-02**	**4.703e-02**	**0.09610.**
Family.History	-2.770e-03	4.468e-02	0.95056
Smoking	-9.477e-02	9.579e-02	0.32248
Obesity	-5.706e-02	4.467e-02	0.20148
Alcohol.Consumption	-6.027e-02	4.549e-02	0.18519
Exercise.Hours.Per.Week	3.988e-03	3.861e-03	0.30158
Previous.Heart.Problems	4.774e-03	4.467e-02	0.91488
Medication.Use	1.064e-02	4.467e-02	0.81166
Stress.Level	-3.205e-03	7.817e-03	0.68177
Sedentary.Hours.Per.Day	-3.849e-03	6.442e-03	0.55013
Income	2.410e-07	2.771e-07	0.38452
BMI	-1.426e-04	3.534e-03	0.96781
Triglycerides	8.971e-05	9.983e-05	0.36887
Physical.Activity.Days.Per.Week	-4.398e-03	9.782e-03	0.65298
Sleep.Hours.Per.Day	**-1.876e-02**	**1.124e-02**	**0.09502.**
Hemisphere	5.487e-02	4.680e-02	0.24108
High.Blood.Pressure	**1.490e-03**	**8.474e-04**	**0.07877.**
Low.Blood.Pressure	-1.171e-03	1.522e-03	0.44178
Unhealthy.Diet	2.321e-02	5.499e-02	0.67297
Healthy.Diet	5.331e-02	5.455e-02	0.32842

Hemisphere-wise Logistic 模型的结果表明，模型整体具有统计显著性（$p < 0.05$），其中截距项为-7.685e-01，具有显著性。在变量方面，Cholesterol（胆固醇）呈现 0.07003 的趋势性显著性，Diabetes（糖尿病）和 Sleep.Hours.Per.Day（每天睡眠小时数）也呈现 0.09610 和 0.09502 的趋势性显著性，而 High.Blood.Pressure（高血压）趋势上显著（$p = 0.07877$）。这表明在半球范围内，胆固醇、糖尿病、睡眠时长和高血压等因素可能对心血管健康有显著影响。

表 3: Logistics 结构模型：Continent-wise

变量	估计值	标准误差	p 值
(Intercept)	-7.716e-01	2.956e-01	0.00904**
Age	1.132e-03	1.177e-03	0.33625
Sex	4.199e-02	5.847e-02	0.47262
Cholesterol	**5.005e-04**	**2.767e-04**	**0.07046.**
Heart.Rate	-4.945e-04	1.087e-03	0.64923
Diabetes	**8.063e-02**	**4.707e-02**	**0.08673.**
Family.History	-4.333e-03	4.471e-02	0.92281
Smoking	-9.039e-02	9.587e-02	0.34572
Obesity	-5.800e-02	4.469e-02	0.19431
Alcohol.Consumption	-6.023e-02	4.552e-02	0.18575
Exercise.Hours.Per.Week	3.960e-03	3.863e-03	0.30534
Previous.Heart.Problems	3.164e-03	4.471e-02	0.94357
Medication.Use	1.262e-02	4.470e-02	0.77760
Stress.Level	-3.579e-03	7.822e-03	0.64727
Sedentary.Hours.Per.Day	-4.109e-03	6.450e-03	0.52413
Income	2.408e-07	2.772e-07	0.38498
BMI	-2.064e-04	3.535e-03	0.95344
Triglycerides	9.022e-05	9.987e-05	0.36628
Physical.Activity.Days.Per.Week	-4.086e-03	9.787e-03	0.67635
Sleep.Hours.Per.Day	**-1.924e-02**	**1.125e-02**	**0.08707.**
Hemisphere	**1.259e-01**	**6.489e-02**	**0.05229.**
High.Blood.Pressure	**1.482e-03**	**8.478e-04**	**0.08048.**
Low.Blood.Pressure	-1.167e-03	1.523e-03	0.44348
Unhealthy.Diet	2.395e-02	5.504e-02	0.66348
Healthy.Diet	5.425e-02	5.461e-02	0.32050
Africa	3.265e-02	1.043e-01	0.75438
Asia	-9.618e-02	8.187e-02	0.24007
Australia	6.584e-02	1.188e-01	0.57949
Europe	-8.572e-02	8.693e-02	0.32410
SouthAmerica	4.428e-02	1.005e-01	0.65949

Continent-wise Logistic 模型显示整体统计显著性（$p < 0.05$），并在变量方面保持了与 Hemisphere-wise 模型相似的趋势。值得注意的是，新增的 Continent（大陆）变量在 0.05229 的 p 值下呈现趋势性显著性，表明大陆水平上的地理位置也可能对模型的结果产生显著影响。

表 4: Logistics 结构模型：Country-wise

变量	估计值	标准误差	p 值
(Intercept)	-5.740e-01	2.985e-01	0.0545.
Age	1.132e-03	1.179e-03	0.3371
Sex	4.326e-02	5.855e-02	0.4600
Cholesterol	**4.839e-04**	**2.770e-04**	**0.0807.**
Heart.Rate	-4.591e-04	1.089e-03	0.6732
Diabetes	**8.057e-02**	**4.713e-02**	**0.0873.**
Family.History	-7.844e-03	4.478e-02	0.8609
Smoking	-8.983e-02	9.598e-02	0.3493
Obesity	-5.756e-02	4.474e-02	0.1982
Alcohol.Consumption	-5.974e-02	4.559e-02	0.1901
Exercise.Hours.Per.Week	4.106e-03	3.868e-03	0.2884
Previous.Heart.Problems	5.329e-03	4.479e-02	0.9053
Medication.Use	1.249e-02	4.476e-02	0.7801
Stress.Level	-3.942e-03	7.836e-03	0.6149
Sedentary.Hours.Per.Day	-4.204e-03	6.458e-03	0.5151
Income	2.551e-07	2.776e-07	0.3582
BMI	-4.532e-04	3.540e-03	0.8981
Triglycerides	8.317e-05	1.000e-04	0.4057
Physical.Activity.Days.Per.Week	-3.644e-03	9.804e-03	0.7101
Sleep.Hours.Per.Day	**-1.947e-02**	**1.126e-02**	**0.0838.**
High.Blood.Pressure	**1.506e-03**	**8.492e-04**	**0.0761.**
Low.Blood.Pressure	-1.114e-03	1.525e-03	0.4652
Unhealthy.Diet	2.904e-02	5.512e-02	0.5983
Healthy.Diet	5.876e-02	5.470e-02	0.2827
Argentina	-1.005e-01	1.385e-01	0.4680
Australia1	-7.542e-02	1.399e-01	0.5899
Brazil	-1.752e-01	1.399e-01	0.2103
Canada	-1.439e-01	1.412e-01	0.3082
China	-1.626e-01	1.417e-01	0.2511
Colombia	-6.681e-02	1.414e-01	0.6366
France	-1.812e-01	1.411e-01	0.1990
Germany	-1.392e-01	1.384e-01	0.3145
India	**-3.447e-01**	**1.462e-01**	**0.0184***
Italy	**-3.512e-01**	**1.443e-01**	**0.0149***
Japan	**-2.732e-01**	**1.431e-01**	**0.0562.**
NewZealand	-1.941e-01	1.423e-01	0.1726
Nigeria	1.925e-02	1.392e-01	0.8900
SouthAfrica	-2.324e-01	1.435e-01	0.1053
SouthKorea	2.985e-02	1.424e-01	0.8340
Spain	-1.941e-01	1.425e-01	0.1732
Thailand	-7.727e-02	1.414e-01	0.5848
UK	-1.823e-01	1.403e-01	0.1939
Vietnam	-1.972e-01	1.429e-01	0.1676

在 Country-wise Logistic 模型中，模型整体具有统计显著性。变量 Cholesterol、Diabetes、Sleep.Hours.Per.Day 等仍保持趋势性显著性。值得关注的是，新增的一些国家级别的变量，如 India、Italy、Japan 等，呈现出对心血管健康的显著性影响（$p < 0.05$）。这表明在具体国家的层面上，文化、饮食和生活方式等因素可能在心血管疾病的发生中起到关键作用。

4 结论与建议

4.1 模型结论

本文使用了多种不同方法的机器学习模型，我们发现总体准确度最好只能达到约 64%，在考虑平衡化样本后，总体准确率达到约 70%。同时，使用复杂模型比简单模型在预测准确度和 AUC 值的表现上均有所提升。针对预测准确率相对不高的问题，我们考虑这与心脏病的特点有关。心脏病发病具有一定的随机性与突然性，良好的作息习惯和体检指标仅能在概率上降低心脏病风险，但不能排除偶然因素的影响，有时短时间内生活习惯上的剧烈变化也会直接导致心脏病风险上升。

在深度学习模型的探索中，我们提出了三个模型，分别为不带 SMOTE 和 DROPOUT 的 10 层神经网络模型，带 SMOTE 不带 DROPOUT 的 10 层神经网络模型，以及 SMOTE+0.1DROPOUT 的 10 层神经网络模型。每个模型的结构和效果通过相应的图示呈现，准确度均为百分之五六十，没有达到很高。

倾向得分回归方法的因果探索中，我们最终识别得到“是否肥胖”和“是否患糖尿病”两个分类变量对心脏病风险有显著影响。而三个 Logistic 回归模型的综合结果揭示了心血管健康影响因素的多层次差异。在全球范围内，Cholesterol（胆固醇）、Diabetes（糖尿病）和 Sleep.Hours.Per.Day（每天睡眠小时数）在半球、大洲和国家层面上均呈现出趋势性显著性，强调了它们对心血管健康的全球性影响。半球层面的模型未显示地理位置变量的显著性，然而，在大洲和国家层面，引入了 Continent 变量和特定国家变量，突显了地理、文化和国家因素对心脏病风险的显著差异。特定国家（如印度、意大利、日本）在各个模型中均呈现出对心血管健康的显著性影响，可能反映了其独特的生活方式和文化特征。

总体而言，本研究通过多角度的机器学习和深度学习模型的应用，深入探讨了心脏病预测的复杂性。尽管准确度仍有提升的空间，但我们的研究为理解心血管健康影响因素提供了深刻的洞察。未来的研究可以进一步探索不同数据特征的影响，以优化模型性能，提高心脏病预测的准确性。

4.2 对心脏病风险预防的建议

经过我们的模型分析，为降低心脏病风险，我们提出以下针对性建议：

关注地域差异：研究不同半球和大洲的地理特征对心血管健康的潜在影响。根据模型结果，为不同地区量身定制预防措施，或许能更有效地降低心脏病风险。

倡导健康睡眠：根据模型结果，睡眠时长与心血管健康关系密切。推动大众建立良好的睡眠习惯，提高睡眠质量，对降低心脏病风险具有积极意义。

有效管理胆固醇和糖尿病：胆固醇和糖尿病在模型中显示出对心血管健康的重要影响。通过调整饮食和生活方式，以及定期监测胆固醇水平和管理糖尿病，能有效降低心脏病风险。

制定国家级的定制化策略：针对模型中显著的国家变量，如印度、意大利、日本等，制定符合各国文化和医疗资源特点的心脏病预防计划。

实施综合性干预：在制定预防措施时，综合考虑年龄、性别、生活习惯和地理位置等多方面因素。通过综合性的干预计划，更全面地降低心脏病风险。

这些建议不仅基于模型分析的统计显著性，还结合了实际效果和领域知识，以确保提供全面、切实可行的心脏病预防方案。在实施过程中，建议与医疗专家紧密合作，确保策略与当地文化和医疗环境相契合。