



**数据挖掘课程设计**

**心脏病风险预测与分析**

|  |  |
| --- | --- |
| 班 级： | 192204 |
| 队 名： | HeartDance |
| 学 号： | 19220432，19220448，19220423，19220446 |
| 姓 名： | 陆昊宇，董文杰，符皓宇，李世博 |

2025年2月28日

# 

目录

[心脏病风险预测与分析 1](#_Toc16835)

[1 项目背景 1](#_Toc9082)

[2 心脏病风险预测 1](#_Toc1158)

[2.1 数据介绍 1](#_Toc20541)

[2.2 算法介绍 2](#_Toc23604)

[2.3 特征构造 3](#_Toc30692)

[2.4 实验结果 6](#_Toc32583)

[3 心脏病相关的预防措施 7](#_Toc20024)

[3.1 XXXX（预防措施一） 7](#_Toc5323)

[3.2 XXXX（预防措施二） 7](#_Toc16082)

[3.3 XXXX（预防措施三）…… 7](#_Toc16299)

[4 结论 7](#_Toc6083)

[5 进度安排 7](#_Toc14660)

[6 团队特色 7](#_Toc27223)

[参考文献 7](#_Toc7557)

## 1 项目背景

随着现代生活节奏的加快和生活方式的改变，心脏病已成为全球范围内威胁人类健康的主要疾病之一。其发病率高、死亡率高、并发症多等特点，使得心脏病的预防、诊断和治疗成为医学研究的重要领域。

本课程设计旨在利用数据挖掘技术，对心脏病进行预测与分析。通过数据预处理、特征构造、模型对比与消融等实验，最终我们发现集成学习策略在该数据集上有良好的拟合效果与泛化性能，同时相较于其他模型，集成学习中的梯度提升模型LightGBM训练速度快、测试精度高，有极佳的应用价值。

## 2 心脏病风险预测

### 2.1 数据介绍

介绍选择了哪些属性，为什么选择这些属性。

数据集中除了Locate标称属性以外的所有属性可以分为三大类，分别为二值属性、序数属性和数值属性。考虑到最终的模型需要面向大众，我们首先去除了Locate属性。而数值属性中身高和体重均是IBM的组成部分，因此我们只保留了IBM。

对于二值属性与序数属性，我们可视化了每一个二值属性与序数属性的心脏病患病情况。

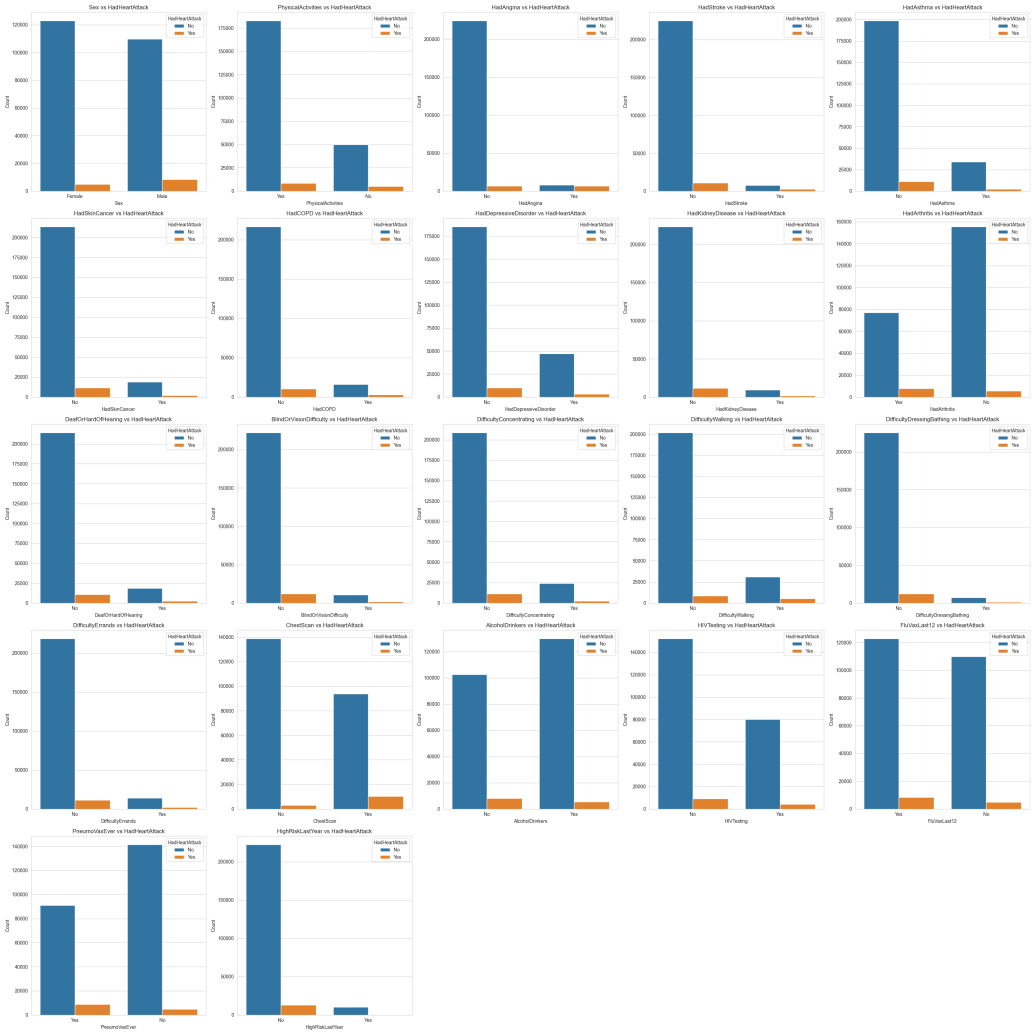


图1 二值属性患病情况

[](..\asserts\multi_cls.png)

图2 序数属性患病情况

如上两图所示，可以看出患有心脏病的属性分布与未患有心脏病的属性分布都是显著不同的，因此我们保留所有的二值属性和序数属性。

### 2.2 算法介绍

由于本次课程设计的任务是二分类，因此我们引入了诸如：线性模型、决策树模型、贝叶斯模型、集成学习模型、懒惰学习模型、深度模型等众多二分类模型进行横向对比实验。

最终发现集成学习模型中的梯度提升决策树LightGBM模型训练速度最快、拟合效果最好，因此我们选择该模型进行进一步的纵向消融实验。

LightGBM模型是一种优化过的梯度提升模型[[[1]](#endnote-0)]，其引入了基于梯度的单边采样（GOSS）技术和独占特征捆绑（EFB）技术，使得该模型在多个公共数据集上训练速度提升了20倍以上同时保持了几乎相同的准确率。如下两表所示：

表1 训练时间对比

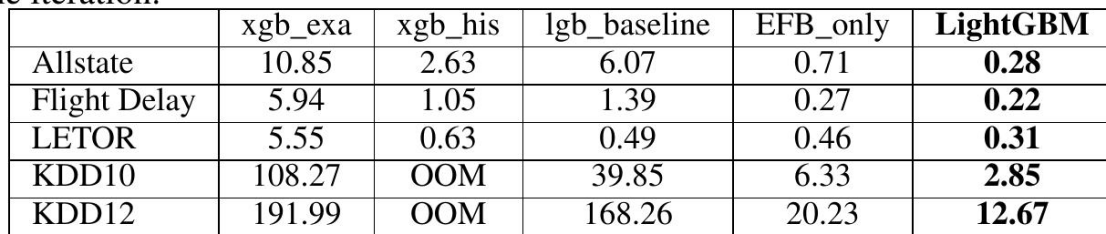
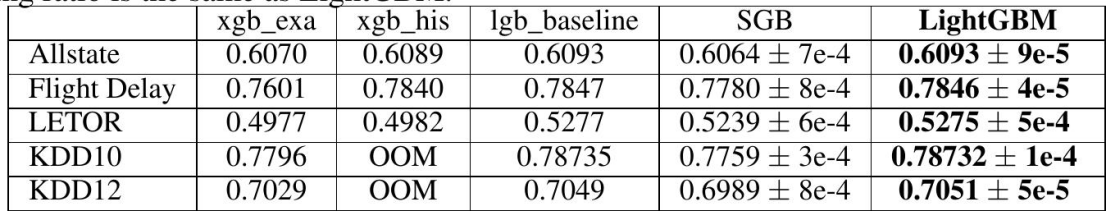


表2 测试AUC对比



可以看出LightGBM非常适用于当前心脏病预测场景。我们首先简单介绍一下梯度提升决策树模型，然后再介绍LightGBM引入的GOSS和EFB两项新技术。

对于梯度提升决策树（GBDT）。它是一种基于集成学习的监督学习方法，属于Boosting算法族的一种。GBDT通过构建多个决策树模型来逐步改进预测结果，每一棵新树的训练目标是减少前一棵树预测误差。具体地，GBDT采用逐步优化的方式，通过计算每个样本的残差（预测误差），并将新树拟合这些残差，从而不断提高模型的预测精度。在训练过程中，每一棵树都会在之前树的基础上进行加权组合，最终的预测结果是所有树的加权平均值。

对于基于梯度的单边采样（GOSS）技术。它是微软团队提出的数据样的一种策略，相比于直接丢弃梯度较小的实例，GOSS 保留所有具有较大梯度值的实例，并对具有较小梯度值的实例进行随机采样。具体地，GOSS首先根据实例梯度的绝对值对数据实例进行排序，并选择前a×100%的实例。然后，它从剩余数据中随机采样b×100%的实例。之后GOSS在计算信息增益时，将采样的具有较小梯度值的数据放大1−b/a倍。通过这样做，模型能够在更加关注训练不足的实例的同时不会大幅改变原始数据分布。

对于独占特征捆绑（EFB）技术。它是微软团队提出的有效减少特征数量的新方法。具体地，在稀疏特征空间中许多特征是互斥的，即它们永远不会同时取非零值。通过精心设计的特征扫描算法，可以将互斥特征捆绑到一个特征中,这样直方图构建的复杂度就会大幅降低，然后就可以显著地加速GBDT的训练而不会损害精度。

### 2.3 特征构造

**缺失值处理**。在使用pandas库加载数据后，使用Frame.info方法发现本次使用的数据集并没有缺失值。

**属性编码**。对于二值属性，我们将所有的No编码为0，Yes编码为1。对于所有的序数属性，我们根据其实际意义进行排序并按升序进行编码。例如在下面的有序字典中，我们将'GeneralHealth'属性的['Poor', 'Fair', 'Good', 'Very good', 'Excellent']依次编码为0,1,2,3,4。

表2 序数属性编码表

|  |
| --- |
| text2number = {  'Sex' : ['Male', 'Female'],  'GeneralHealth' : ['Poor', 'Fair', 'Good', 'Very good', 'Excellent'],  'LastCheckupTime' : ['Within past year (anytime less than 12 months ago)', 'Within past 2 years (1 year but less than 2 years ago)', 'Within past 5 years (2 years but less than 5 years ago)', '5 or more years ago'],  'RemovedTeeth' : ['None of them', '1 to 5', '6 or more, but not all', 'All'],  'HadDiabetes' : ['No', 'No, pre-diabetes or borderline diabetes', 'Yes, but only during pregnancy (female)', 'Yes'],  'SmokerStatus' : ['Never smoked', 'Former smoker', 'Current smoker - now smokes some days', 'Current smoker - now smokes every day'],  'ECigaretteUsage' : ['Never used e-cigarettes in my entire life', 'Use them some days', 'Use them every day', 'Not at all (right now)'],  'AgeCategory' : ["Age 18 to 24", "Age 25 to 29", "Age 30 to 34", "Age 35 to 39", "Age 40 to 44", "Age 45 to 49", "Age 50 to 54", "Age 55 to 59", "Age 60 to 64", "Age 65 to 69", "Age 70 to 74", "Age 75 to 79", "Age 80 or older"],  'CovidPos' : ['No', 'Tested positive using home test without a health professional', 'Yes'],  'TetanusLast10Tdap': ['Yes, received Tdap', 'Yes, received tetanus shot but not sure what type', 'No, did not receive any tetanus shot in the past 10 years', 'Yes, received tetanus shot, but not Tdap']} |

**异常值处理**。在对属性编码后，我们对所有属性计算了IQR指数并利用分箱法寻找可能的异常值。

**规范化**。为了避免不同属性的量纲差异导致的系统误差，我们对所有编码后的属性进行min-max归一化操作。

**数据划分**。规范化好以后，我们组将数据集按照6:2:2的比例划分为训练集、验证集和测试集。其中验证集当且仅当算法没有集成交叉验证时使用。

**数据平衡**。通过2.1中的可视化结果可以明显看出本数据集存在严重的正负样例失衡的情况，因此我们组在数据划分好以后，对训练集进行了数据平衡工作。考虑到正例极少，降采样会使得数据规模过小从而导致模型欠拟合，因此我们组采用了SMOTE算法对训练集进行过采样。该算法的过采样逻辑主要以插值的方式实现，相比于多次重复采样提升了数据的多样性，为后续模型的高泛化性做好铺垫。平衡结果如下两图所示：



图3 平衡后的二值属性患病情况



图4 平衡后的序数属性患病情况

### 2.4 实验结果

**对比实验部分**。我们分别使用了线性模型、决策树模型、贝叶斯模型、集成学习、懒惰学习、深度模型进行二分类模型的对比实验。针对当前心脏病预防检测的场景，我们将召回率作为第一指标，而为了防止模型过拟合，我们辅以准确率与ROC曲线下的面积AUC作为第二、三指标。下表展示了上述模型在测试集上的预测情况：

表3 对比实验结果表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | 准确率 | 召回率 | AUC |
| 线性模型 | 逻辑回归 | 0.8179 | 0.7659 | 0.8646 |
| 决策树模型 | CART | 0.9180 | 0.2493 | 0.6030 |
| 贝叶斯模型 | 高斯贝叶斯 | 0.7965 | 0.7607 | 0.8531 |
| 多项式贝叶斯 | 0.8039 | 0.7592 | 0.8576 |
| 伯努利贝叶斯 | 0.7912 | 0.7514 | 0.8523 |
| 集成学习 | **LightGBM** | 0.8228 | **0.7667** | **0.8843** |
| Adaboost | 0.7988 | 0.7544 | 0.8693 |
| 随机森林 | **0.9465** | 0.3003 | 0.8670 |
| 懒惰学习 | K近邻 | 0.8812 | 0.4592 | 0.7206 |
| 深度模型 | FNN | 0.8158 | 0.7324 | 0.7765 |

**消融实验部分**。从对比实验结果可以看出，集成学习策略在当前数据集上的拟合效果较好。

## 3 心脏病相关的预防措施

### 3.1 XXXX（预防措施一）

给出预防措施的描述，以及实验和数据的支撑证据。

### 3.2 XXXX（预防措施二）

给出预防措施的描述，以及实验和数据的支撑证据。

### 3.3 XXXX（预防措施三）……

给出预防措施的描述，以及实验和数据的支撑证据。

## 4 结论

从项目的角度总结本次课设的主要成果。

## 5 进度安排

介绍项目的分解情况，以及每个队员的任务分配以及完成情况。

## 6 团队特色

介绍本次课设团队最具有特色的地方。

## 参考文献

如果使用了参考文献，附加在这一章节。

1. [] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu. "[LightGBM: A Highly Efficient Gradient Boosting Decision Tree](https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html)". Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 3149-3157. [↑](#endnote-ref-0)