華東師範大學
East China Normal University

统计机器学习课程论文

基于 stacking 融合模型的老年人健康状况预测

姓　　名： 吴欣　　秦佳玥　　万翊臻

学　　号： 10205000506　10205000512　10205000467

学　　院： 经管书院

专　　业： 统计学

2023 年 6 月

（注：姓名先后不代表贡献程度，本小组 3 人对小组作业贡献平等）

摘要

老年人健康问题一直是社会关注的焦点。随着人口老龄化的加速，大量老年人的健康状况也日益受到社会各方的关注与重视。针对 ICR 公司提出的简单机器学习模型难以准确预测老年人患病情况的问题，本小组建立了一种 Stacking 融合模型给出了有效的解决。首先，在对比赛给出的数据集进行可视化分析与探索后，本文对原始数据进行填补、降维等预处理尝试，并根据预测目标选取 balanced log loss 作为分类性能度量指标。其次，本文根据数据特性，选取 CatBoost、LightGBM、随机森林、支持向量机、XGBoost 五个单模型，在对其分别基于 Optuna 进行超参数优化后，综合考虑训练时间与 balanced log loss，选出 lightGBM 作为最优单模型，并对模型进行解释，分析影响预测患病概率的变量。接着，本文通过对各种备选组合的比较，选出 CatBoost 和 XGBoost 作为 Stacking 方法的基学习器，具有较好泛化能力的 Random Forest 作为 Stacking 方法的元学习器，成功建立最终的 Stacking 融合模型，该模型取得了 0.405 的 balanced log loss 值。

关键词：老年人健康，分类预测，stacking 融合模型

一、选题介绍

(一) 选题动机

随着全球人口老龄化的持续加剧，老年人的健康问题受到社会各方的广泛关注。随着年龄的增加，老年人的身体机能和免疫能力都逐渐衰弱，更容易罹患慢性基础疾病，如高血压、糖尿病、心脏病等，甚至诱发并发症。这些需要长期进行药物治疗的疾病，不仅会给老年人带来身体上的痛苦，同时也给社会经济和公共医疗体系带来沉重的负担。

针对老年人健康问题，最新的生物信息学和机器学习技术等科学手段可以有效评估、预测老年人的健康状况和寿命，及时发现疾病并给予有效治疗。这可以较好地关怀老年人的健康状况，提高老年人健康水平，同时，还可以提升医疗资源的使用效率，缓解医务人员和机构的压力，进而实现优化医疗资源配置，构建更加公平、有效的医疗制度，为社会和谐发展的未来奠定坚实的基础。

(二) 问题介绍

本次比赛的主办方 InVitro Cell Research, LLC (ICR)，要求参赛者根据不同的健康特征数据，预测老年人是否患有三种常见的疾病：心脏病、高血压和糖尿病。传统的医学手段往往需要漫长而侵入性的过程来确定某人是否患有这些疾病。使用预测模型可以缩短此过程，并通过收集与条件相关的关键特征，并对这些特征进行编码，从而对患者的详细信息进行保密，提高数据安全性。

虽然心脏病、高血压和糖尿病在老年人中比较常见，但依然需要采集大量健康数据并进行分析才能进行准确的诊断。目前医疗领域存在数据质量、模型精度以及隐私保护等问题，如何运用最新的机器学习技术构建精准、可靠的预测模型，并基于脱敏数据进行有效预测，是当前医疗领域需要解决的重大问题，也是本次比赛的核心焦点。

本次比赛的目标是预测一个老年人是否患有三种疾病中的任何一种。“1”类表示该老年人患有三种疾病中的任何一种或多种，“0”类表示该老年人没有罹患三种疾病中的任何一种。参赛者需要创建一个针对健康特征测量值进行训练的模型，探索某些健康特征的测量与潜在患者状况之间的关系。

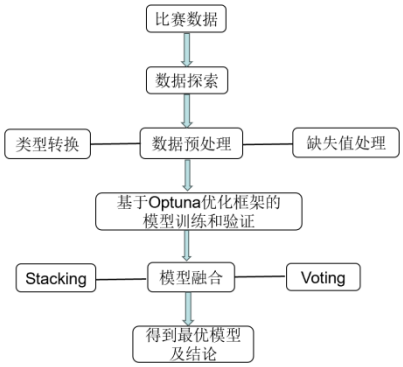
(三)章节安排

首先，我们对竞赛给出的数据集进行基本的了解，包括数据来源和各个数据集的内容与变量含义，在此基础上确定了建模全程的训练集与测试集的划分，这部分见第二章。

接着，我们重点对train数据集的各个特征进行了可视化分析和数据预处理（包括缺失值，异常值处理和降维尝试），并根据预测目标选定分类性能度量指标，这部分见第三章。

在上述分析的基础上，我们开始正式建立模型，首先从初步挑选的CatBoost、LightGBM、随机森林、支持向量机、XGBoost五个单模型中选出最优单模型，接着通过尝试不同模型组合，选出CatBoost和XGBoost作为Stacking方法的基学习器，具有较好泛化能力的Random Forest作为Stacking方法的元学习器，并通过分析变量重要性，给出预测最优Stacking方法模型解释，这部分见第四章。

最后，我们评价最终模型预测效果，并对此次研究进行了回顾反思，总结出几点教训与未来展望，这部分见第五章。



二、数据介绍

(一)数据来源

本论文所使用的数据集来源于InVitro Cell Research, LLC (ICR)，是由ICR的科学家们采集的老年人健康数据。ICR是一家专注于再生和个性化预防医药的私人资助公司。

该数据集可以从竞赛网站（https://www.kaggle.com/competitions/icr-identify-age-related-conditions/data）上“ICR-Identifying Age-Related Conditions”这一比赛页面上查看，报名参赛后可下赛全部数据集。

(二)数据集介绍

该比赛一共提供了4个数据集，分别为train数据集，test数据集，greeks数据集和sample_submission数据集。

train数据集是比赛提供的训练集，共有617个带有缺失值的样本，涵盖了来自不同地区和族群的老年人信息。该数据集包含56个匿名化的健康特征和一个目标变量Class，56个健康特征中除了EJ这一变量是分类特征外，其余都是数值型特征，Class的取值为0或1，0表示未被诊断出三种疾病中的任何一种，1表示被诊断出三种疾病中的一种或多种。

test是比赛提供的测试集，其指明最终目标是预测一个样本属于两个类别（0或1）的概率。

greeks数据集仅适用于训练集，其中的Alpha列用于表示是否存在相关疾病，A表示没有诊断出相关病症，对应于Class=0；B，D，G分别表示被诊断出其中的一种病症，对应于Class=1。此外，该数据集还包含Beta、Gamma和Delta三个实验特征以及Epsilon这一日期特

征（表示数据采集的时间），根据比赛声明，测试集中的所有数据都是在训练集采集之后进行收集。

sample_submission数据集是一个示例提交文件，展示了提交文件的正确结构和格式要求。

(三)训练集和测试集的构造

train数据集共含有617个样本，按7:3的比例划分为训练集1、验证集1。其中训练集1包含431个样本，验证集1包含186个样本。为确保数据集的随机性和分布，以上数据集的分割是在随机的情况下产生的，保证了训练集、验证集和测试集之间特征分布的一致性。同时，为了结果的可复现性，设定随机数种子为48。此次划分用于最终选出的模型的效果对比。

接着，我们对训练集1进行了二次划分，选择五折交叉验证。二次划分的目的是为了进行单集成模型的效果对比，Stacking模型元学习器参数调优。分割依然保证了划分随机性和结果可复现性。

最后，为了进行单集成模型的参数调优及Stacking模型基学习器的选择，在训练集2上进行了第三次划分，仍选择了五折交叉验证的形式，示意图如下。

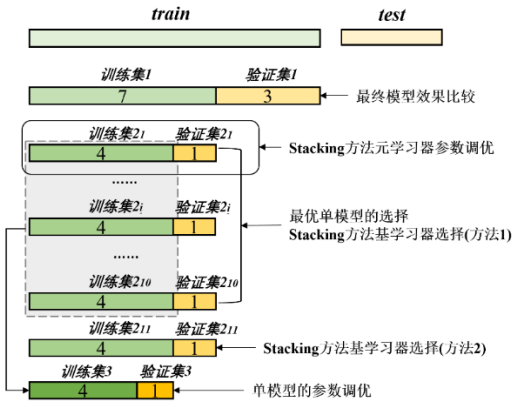


图1

三、数据分析

(一)数据预处理

1. 数据集特征分析

我们首先对train数据集进行可视化处理，图x展示了该数据集中每个特征的唯一值数量，从图中可以看出，train数据集包含617个唯一的Id值。图x的小提琴矩阵图展现了数据集中一个特征的分布情况，包括分位数、中位数，以及数据点分布的密度，GH和GI等特征的大部分分布接近0，但具有很长的尾部，这可能意味着数据集中可能存在异常值，但考虑到这些值与医学相关且具体含义未知，我们并未对其进行异常值处理。此外，从小提琴矩

阵图中可以看出个别特征的分布差异较大，后续可能需要对数据进行降维处理。

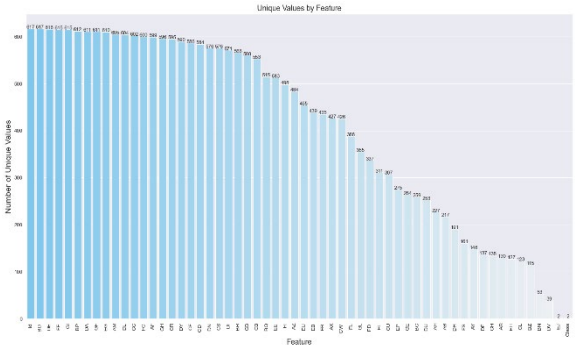


图2

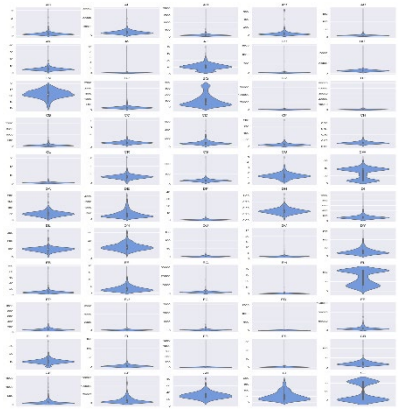


图3

另外，考虑到我们的目标是预测Class的类别，所以我们需要检验是否有显著的类偏差，结果如图X所示，可以看出，数据存在较为严重的变量不平衡问题，在后续模型训练的参数设定部分需要对该问题进行处理。

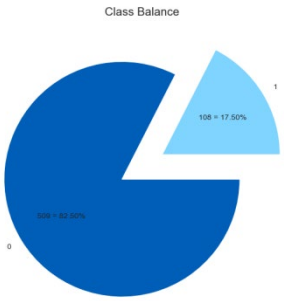


图4

最后，我们对train数据集绘制了一个矩形热图（图5），展示了数据集中各列特征之间的Spearman相关系数。由于特征数量过多，相关系数的数值没有显示出来。然而，它们颜色的深浅代表了相关性强弱。深红色表示强正相关，深蓝色表示强负相关。从图中可以看到，没有一个单一的特征与类目标有很强的相关性，但是个别特征彼此之间有很强的相关性。

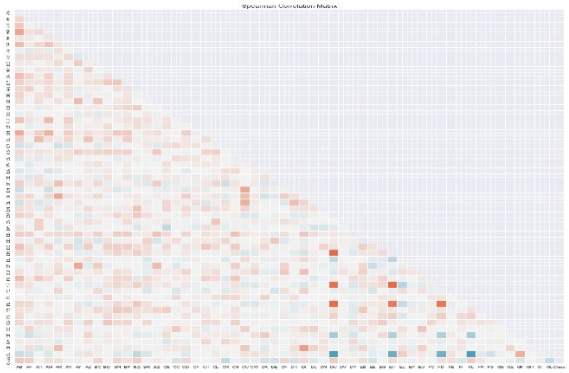


图5

2. 缺失值填补

通过对数据集的简单分析，我们发现train数据集中存在少量数据缺失的问题，这可能与错误、欠慎的调查、测量误差等问题相关。考虑到在上述过程中保留了数据中的异常值，我们选择对缺失值采取中位数填补的方法，该方法具有较强的鲁棒性，不易受到异常值的影响，从而获得较为稳定的结果。

3. 特征降维尝试

由于在数据集特征分析中，我们发现train数据集中极少数特征具有较强的相关性，因此我们对数据进行了特征工程和降维处理的尝试，但没有取得更好的效果。首先，我们尝试了主成分分析（PCA）的降维方法，但降维后所有模型的效果都出现了不同程度的下滑，该结果表明，PCA降维方法并不适合我们的数据集。其次，考虑到数据集中既包含数值型特征，又包含分类特征，我们尝试了随机森林算法进行特征筛选，但由于该方法忽略了数据集中的分类变量，并且在变量重要性上表现得比较平均，也没有取得很好的结果。综上，我们决定在此阶段不对数据进行特征工程和降维处理。

(二)分类性能度量指标

由于我们需要处理的预测问题是二分类问题，并且根据前述分析，数据集存在变量不平衡问题，使用简单的精度或F1值等指标可能会带来误导。因此，为了更好地评估模型在不平衡数据集上的性能，我们选择使用比赛提供的balanced log loss指标，它考虑了二分类问题中不同类别的权重,并且对两个类别之间的样本数量差异进行了平衡。其公式为：

$$\text{balanced log loss} = \frac{\frac{-1}{N_0}\sum_{i=1}^{N_0} y_{0i} log p_{0i} - \frac{-1}{N_1}\sum_{i=1}^{N_1} y_{1i} log p_{1i}}{2}$$

四、模型构建

(一)基于 Optuna 优化框架的单模型构建

1. 参数优化和模型建立

集成学习通过构建并结合多个学习器来完成学习任务，与个体学习器相比，集成学习模型在预测准确性上表现更好。因此，基于模型特点和数据特性，我们选取 CatBoost、LightGBM、随机森林、支持向量机、XGBoost 五个集成学习算法对训练集 1 进行交叉嵌套验证，并比较 5 个单模型的效果。其中，CatBoost 是一种基于梯度提升决策树的框架，具有高效性能、泛化能力、处理类别特征和防止过拟合的能力，适合于处理大规模数据；LightGBM（Light Gradient Boosting Model）是一种基于梯度提升决策树的框架，在大规模数据集上训练速度快、内存占用低，具有较好的准确性和泛化能力；随机森林（Random Forest、RF）是一种集成学习算法，具有在高维数据和大规模数据集上表现良好、具有较好的稳定性的特性；支持向量机（Support Vector Machine，SVM）是一种监督学习算法，适用于二分类和多分类问题，并具有高效处理大规模数据集的能力；XGBoost (eXtreme Gradient Boosting) 是经过优化的分布式梯度提升库，在并行计算效率和预测泛化能力上都表现优良。

在确定单模型后，我们设置适当的超参数以提高模型的准确性和性能。为选取合适的参数组合，我们选择 Optuna 这一自动超参数优化框架对模型进行参数调优。Optuna 默认使用 optuna.samplers.TPESampler 方法进行参数优化搜索，这是贝叶斯优化的一种，通过反复计算不同参数值的目标函数值选取最佳参数组合。与网格搜索、随机搜索等参数优化方法相比，Optuna 具有并行的分布式优化、适用于多个机器学习框架、对不理想试验剪枝等多个特性，能够更高效地搜索超参数空间，加速模型调优过程。此外 Optuna 可以通过调用 optuna-dashboard 记录每次学习过程。

Optuna 优化框架的核心为目标函数（objective）、单次试验（trial）和研究（study），其中 objective 负责定义待优化函数并指定参数数范围，trial 对应着 objective 的单次执行，而 study 则负责管理优化，决定优化的方式，记录总试验的次数、试验结果等功能。本文设置目标函数为 balanced log loss，试验次数为 200 次，优化方式为取目标函数的最小值。

为得到每个单模型的最优参数组合和增加统计的显著性，我们进行嵌套交叉验证，在 10 次迭代中划分出的训练集 2 上进行参数优化、选择出最优参数组合，同时记录单模型在训练

集 2 上的训练时间和参数优化过程，作为评估单个模型效果的指标之一。完成模型的训练后，计算验证集 2 上的 balanced log loss，作为评估单个模型效果的另一个指标。

以 LighGBM 为例，其所选调优参数、调优范围及最优参数见表 1，参数优化过程见图 6。

表 1 LightGBM 所选调优参数、调优范围及最优参数

参数名称	参数含义	范围	最优参数
n_estimators	boosting 的迭代次数	(1000,30000)整数	22976
learning_rate	学习率，用于减少梯度步长	(0.01,0.3)小数	0.294
num_leaves	一棵树上的叶子节点个数	(10,3000)间隔 20 的整数	2870
max_depth	限制树模型的最大深度，用于防止过拟合	(3,12)整数	4
reg_alpha	L1 正则化参数，用于控制模型复杂度	(0.01,0.7)小数	0.515
reg_lambda	L2 正则化参数，用于控制模型复杂度	(0.01,0.7)小数	0.265
bagging_fraction	不进行重采样的情况下随机选择的数据比例，用于防止过拟合，加速训练	(0.2,0.95)间隔 0.1 的小数	0.9
bagging_freq	bagging 的频率，用于启动 bagging	1	1
feature_fraction	每次迭代中随机选择部分特征，用于加速训练，防止过拟合	(0.2,0.95)间隔 0.1 的小数	0.600
min_child_samples	一个叶子节点中最小的数据量，用于防止过拟合	(10,100)间隔 5 的整数	85
colsample_bytree	每次迭代中随机选择部分特征	(0.1,0.9)小数	0.876
random_state	随机数种子	48	\
categorical_feature	指定分类特征	第 39 列（即”EJ”列）	\
class_weight	类型权重，用于减少木匾变量不平衡对模型效果的影响	'balanced'	\
early_stopping_round	如果一个验证集的度量在 early_stopping_round 循环中没有提升，将停止训练。	200	\

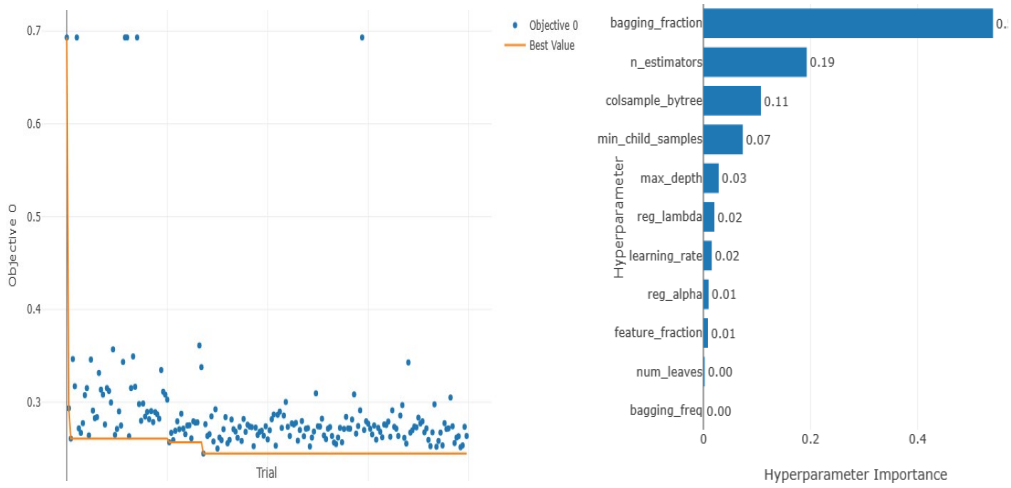


图 6 基于 Optuna 优化框架下的 LightGBM 调参过程

（左图为目标函数值随迭代次数的变化情况，右图为参数重要性）

从图 6 的左图中，我们观察到随着迭代次数的增加，模型的最优目标函数值逐渐收敛到 0.245；右图显示在超参数中 bagging_fraction 相对重要性为 0.541，说明其对模型性能的影响程度最大。

2. 预测效果比较

为筛选出表现效果最佳的单模型，我们将迭代 10 次得到运行时间和 balanced log loss 绘制成箱线图，便于更直观地进行对比。

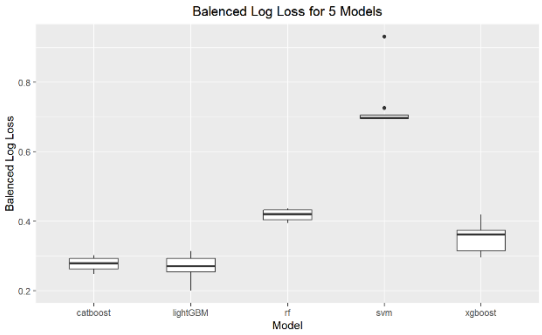


图 7 单模型在验证集 2 上的 balanced log loss

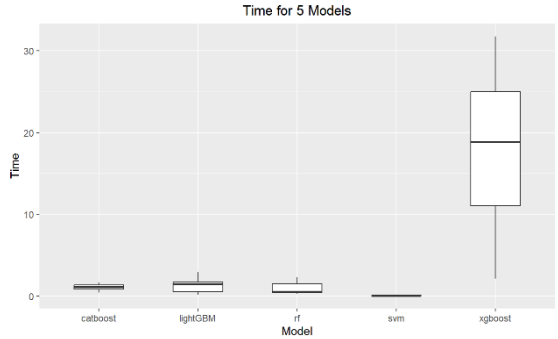


图 8 单模型的训练时间

图 7、8 显示，LightGBM 的预测误差最小，且运行时间较短。虽然 SVM 在运行效率上具有一定优势，但是该模型的预测误差较大。因此，我们最终选择 LightGBM 作为最优单模型，与后续的模型融合方法进行进一步比较。我们将训练集 1 输入到 LightGBM 中进行拟合，将验证集 1 输入到 LightGBM 中进行预测，经过模型融合后得到 balanced log loss 为 0.449。

3. LightGBM 模型效果解释

a) 变量重要性

可解释性是机器学习中不可或缺的部分，我们通过基尼重要性、置换法和 shap 衡量基于 LightGBM 模型的变量重要性。使用 model. feature_importances 计算各变量的“Gini Importance”，发现在每一颗树中，由 DA 变量形成形成的分支节点的 Gini 指数下降程度之和最大，DU 变量次之。使用置换法看待各个变量的重要性，发现 DU 变量依然有最高的重要性，在这类重要性的衡量中，BQ 变量较 DA 更为重要，说明随机排列 BQ 的值会导致模型预测效果更大幅度地下降。shap 变量重要性排序中，DU 和 BQ 是最重要的两个变量。

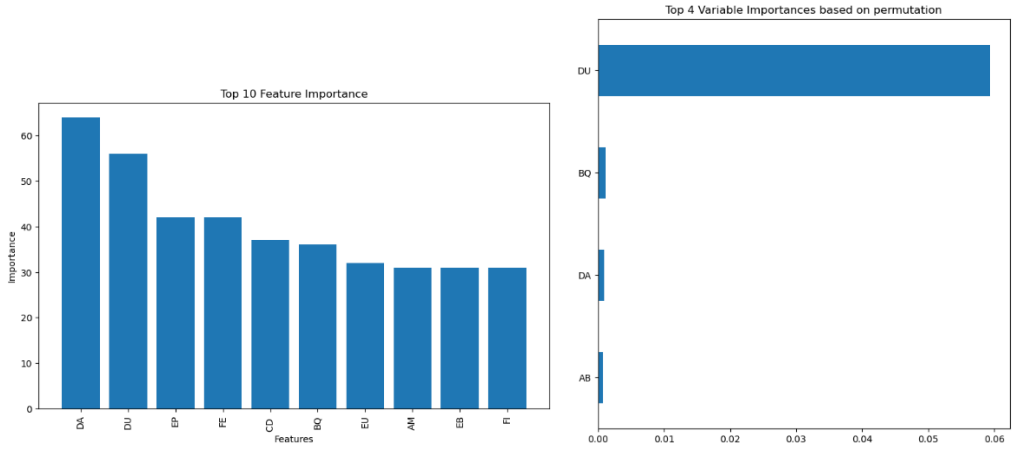


图 9 变量重要性（左图为基尼重要性，右图为置换法）

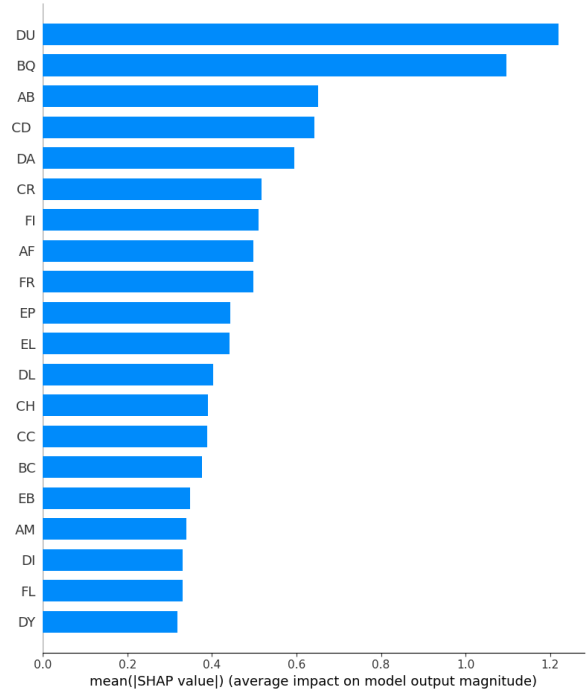


图 10 20 个变量的 SHAP 特征重要性

b） 变量作用效果分析

绘制 DA，BQ 的偏依赖图如图 11，可以看到随着 DA 的预测值的增加，样本患病的概率将降低，增加到一定程度时（约为 60），不再对是否预测为患病产生影响，样本患病的概率将维持不变。而 BQ 值的增加则可能会引起样本预测为患病的概率的增加：在 BQ 值较低的范围内，其增加并不影响患病概率预测，当 BQ 大于某个值时（约为 95），对患病的概率预测将迅速提升，BQ 达到 200 左右时，该指标的增加不会对预测造成影响。因此，这一指标很可能有利于疾病发展进程的监控。

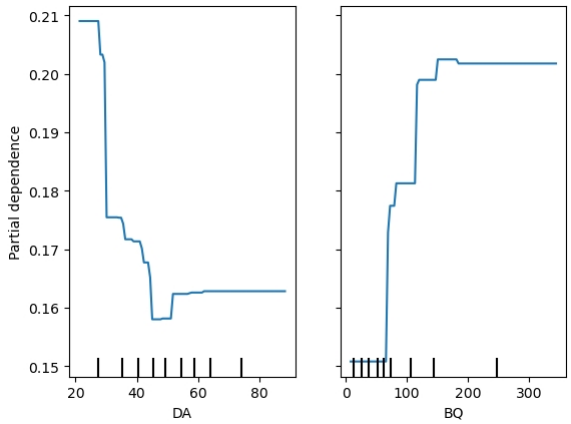


图 11 DA、BQ 的偏依赖图

选取第一个观测样本，查看各个变量对其预测值的影响。根据对官方文档的阅读，在我们的模型中，base value 为患病与否的概率对数值，计算得其值为-4.848，即所有样本的平均为 1（患病）的水平为-4.848。如图 X，在该个体样本中，变量 EP 使得该样本的预测结果从 base value 向负向偏离最大，FR 则使得该样本的预测结果从 base value 向正向偏离最大，其中，BQ 变量使得该样本的预测结果从 base value 向正向偏离，即增大了样本被预测为患病

的概率。在这个观测样本中，有更多的变量使得预测值负向偏离更大幅度，这些变量都与该样本被预测较低的患病概率有关。

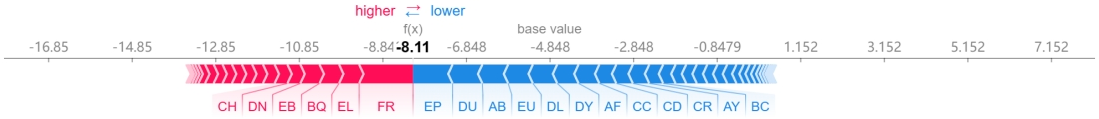


图 12 第一个样本的 force_plot

查看前三个最重要的变量对所有样本预测值的影响，如图 13，可以发现相似的变量呈现了集群性的预测波动，说明 LightGBM 模型相对较为合理。对于大部分样本，三个变量均会统一倾向于使得预测值负向或正向偏离 base value，但导致的偏离程度有所不同。而对于少部分样本，三个变量的表现有一定分歧。

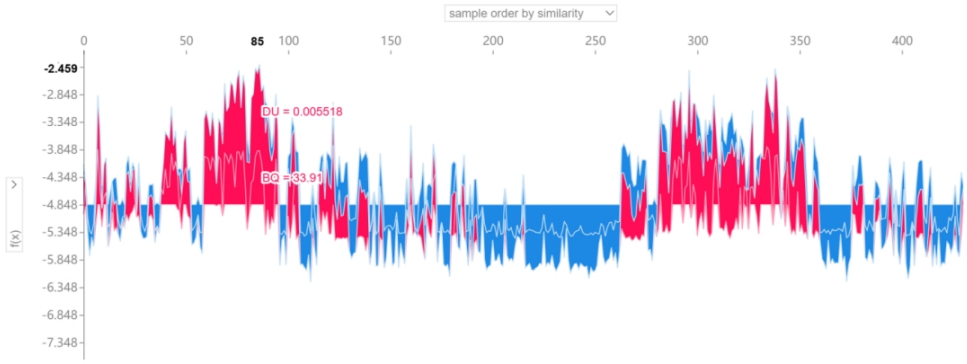


图 13 'DU','BQ','AB' 对所有样本观测预测值的影响

(二) Stacking 模型的构建与优化

为了进一步提升模型预测效果，使用 Stacking 方法集成不同的个体学习器来提升预测效果。选择合适的基学习器对 Stacking 模型的效果十分重要，我们使用两种选择基学习器的方式：一种从单模型学习能力出发，根据前述单模型的训练效果选出最佳的三个单模型，集成预测性能较好的基学习器；另一种从单模型之间的差异度出发，选择差异度较大，预测能力较好的三个模型作为基学习器，综合不同算法抽取特征的优势。元学习器则应选择较为稳健，或相对简单的模型。

1. 基学习器的选择

我们基于以下两种方式进行基学习器的选择。

选择方式一：单模型学习能力

根据单模型训练效果，可以选出在训练集预测效果最好的三个模型，从前述分析结果（图 7）来看，应采用 LightGBM，CatBoost 和 XGBoost 三个单模型作为基学习器。

选择方式二：单模型之间差异度

模型之间的差异度可以由模型预测结果的相关性来衡量（2022）。我们对训练集 1 重新划分，令各模型在训练集 2_{11} 上进行参数调优，在验证集 2_{11} 上进行预测，得到每个模型的预测结果。在训练集 2_{11} 上得到的参数调优结果作为单模型的最优参数保留，作为后续 Stacking 方法的基学习器，在 Stacking 中，不再对基学习器重复进行参数调优。LightGBM 的最优参数如表 1 所示，其余模型的最优参数可见附录。

我们采用皮尔逊相关系数来分析各个模型之间的相关程度，由于 SVM 模型在前述单模型效果比较时表现十分不理想，我们在第二种选择方式中也不将其纳入考虑。

绘制热力图如图 14 所示，较深的颜色表明预测结果的相关性较强，相关系数越接近 1。

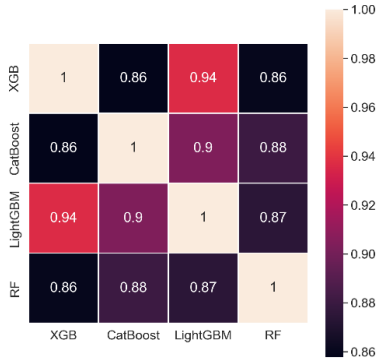


图 14 单模型预测结果热力图

由图 14，LightGBM，XGBoost 和 CatBoost 模型的预测结果呈现了较强的相关性，我们推测这样的原因是三者在建模过程上的相似性：不仅都以梯度提升决策树算法为基础，而且具有相似的优化目标和损失函数。另外，我们使用的数据集样本量较小，可能导致尽管三个模型在实现过程中有细节的差异，也无法在预测结果上呈现较大的差别。Random Forest 与这三个模型的相关性都相对较低，我们认为这样的原因可能是 RandomForest 使用的是另一种基于决策树的算法，同时，它还使用 Bagging 技术（随机抽样）来生成不同训练集，并支持对特征进行抽样。而另外三个模型的运行速度较慢，在考虑到时间不充分的情况下，我们在训练过程中未使用随机抽样的方法，仅基于特征进行采样。因此，Random Forest 模型的预测结果相比其他三个模型，有一定差别。

为了保证 Stacking 方法的效果，首先选择单模型表现最优的 LightGBM 和次优的 CatBoost，再参考热力图，选择与 LightGBM 模型相关性最低的 RandomForest，由这三个模型作为第二种 Stacking 方法的基学习器。

2. 元学习器的选择

Stacking 方法的元学习器需要选择泛化性能较强的学习器，以降低过拟合的风险（2020）。由于 Random Forest 采用 bagging 的融合策略构建多棵决策树，每棵决策树只使用部分数据样本和部分特征，增大模型之间的差异性，相比 Boosting 模型降低了方差和泛化误差，具有更高的泛化性能，我们选择该模型作为尝试的元学习器之一。另外，还采用 Logistic 回归模型作为元学习器，该模型适用于二分类问题，输出值表示样本属于某一类的概率，适用于我们研究的问题，同时相比前述模型，较为简单。两个元学习器均基于训练集 1 使用 optuna 进行参数优化，RandomForest 回归模型所选的优化参数，优化范围及优化结果见表 2。Logistic 回归模型的结果可参考附录。

表 2 Stacking 方法中 Random Forest 作为元学习器，
LGBM, CatBoost, XGBoost 为基学习器的参数调优结果

参数	调优范围	最优值
n_estimators	(100, 1000)	644
criterion	"gini", "entropy"	"gini"
max_depth	(3, 12)	3
min_samples_split	(2, 10)	10
min_samples_leaf	(1,5)	1
max_features	"sqrt", "log2"	"sqrt"

3. Stacking 方法中不同学习器组合方式效果

在验证集 1 上评估不同 Stacking 方法的预测效果，计算 balanced log loss 值及元学习器拟合时间。同时输出元学习器的变量重要性，以查看不同基学习器在集成方法中的重要性。

对于元学习器为 RF 的组合，使用 model.feature_importances_，计算的是各变量的“Gini Importance”，或称平均减少不纯度，表示 RF 的每一棵树中，由某一变量形成的分支节点的 Gini 指数下降程度（或不纯度下降程度）之和。

对于元学习器为 Logistic 回归的组合，使用模型系数（model.coef_）。该系数表示自变量每变化一个单位，预测的发生某事件与否的概率比的对数值的改变。

在检查基学习器为 LGBM，CatBoost 和 RF，元学习器为 RF 的组合的变量重要性时，发现该模型内，RF 对应的重要性为 0，于是只使用 LGBM 和 CatBoost 作为基学习器，RF 为元学习器再次进行 Stacking。五种组合及 LightGBM 单模型的实验结果如表 3：

表 3

基学习器	元学习器	balanced log loss	变量重要性 (值与基学习器一一对应)	拟合时间（秒）
LGBM, CatBoost, XGBoost	RF	0.405	[0.321, 0.354, 0.326]	1.162
LGBM, CatBoost, XGBoost	Logistic 回归	0.446	[0.316, 0.356, 0.328]	0.966
LGBM, CatBoost, RF	RF	0.410	[0.411,0.589,0]	0.002
LGBM, CatBoost	RF	0.407	[0.453,0.547]	0.001
LGBM, CatBoost, RF	Logistic 回归	0.412	[0.389,0.460,0.151]	0.003

从表 3 的结果看到，尽管 LightGBM 在单模型预测效果中表现最优，它在 Stacking 方法里并不占有最高的变量重要性，在所有组合中，CatBoost 均占有最高的重要性，在模型中的贡献最大。在所有组合中，以 LGBM, CatBoost, XGBoost 为基学习器，RF 为元学习器的 Stacking 方法在验证集 1 上的预测效果最好（命名为预测最优 Stacking 方法），具有最低的 balanced log loss（0.405），但拟合时间也最长。以 LGBM,CatBoost 为基学习器，RF 为元学习器的 Stacking 方法的预测效果也较好，仅比最优值低 0.002，而以 Logistic 回归模型作为元学习器的组合相对表现出较弱的预测能力。

将 Stacking 方法与 LightGBM 单模型的效果对比，结果如表 4：

表 4

模型	balanced log loss	拟合时间（秒）
LightGBM	0.449	1.279
LGBM, CatBoost, XGBoost+RF	0.405	1.162
LGBM, CatBoost, XGBoost+Logistic 回归	0.446	0.966
LGBM, CatBoost, RF+RF	0.410	0.002
LGBM, CatBoost+RF	0.407	0.001
LGBM, CatBoost, RF+Logistic 回归	0.412	0.003

由此可发现相比 LightGBM 单模型，Stacking 方法都在验证集 1 上表现出了预测能力的提升，虽然所有 Stacking 方法的元学习器拟合时间都短于 LightGBM 单模型，但这样的原因是因为我们只统计了元学习器拟合时间来代表 Stacking 方法的拟合时间，而未能考虑基学习器的预测时间，故而 LightGBM 单模型是否在预测时长上表现也较差有待商榷。

4. 预测最优 Stacking 方法模型解释

a) 变量重要性衡量

查看除了基尼重要性以外的变量重要性衡量方法，例如置换法，如图 15，LightGBM 的置换法重要性最高，说明随机排列其值会导致模型预测效果的大幅下降。这与 LightGBM 作为单模型的预测效果最佳是一致的。由于我们的预测变量都有较高相关性，置换其中一个变量的值后，其他正确变量会继续对模型进行贡献，导致被置换变量的重要性被低估，这可能是 XGBoost 的置换法重要性较低的原因。另一种方法是比较 shap 变量重要性，该重要性集成了每个样本的该特征重要性之和。如图 16，CatBoost 具有最高的 shap 变量重要性，LightGBM 次之，三个模型对预测效果的影响都是正向的。

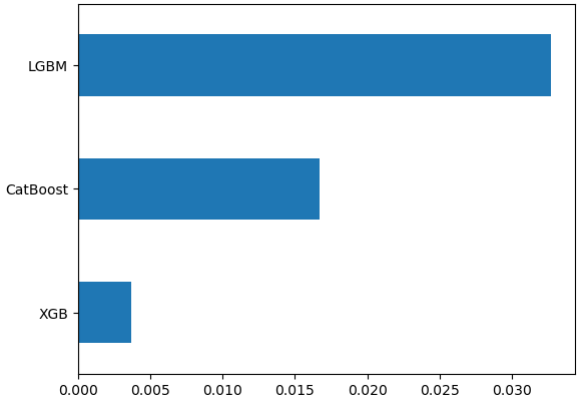


图 15 置换法变量重要性

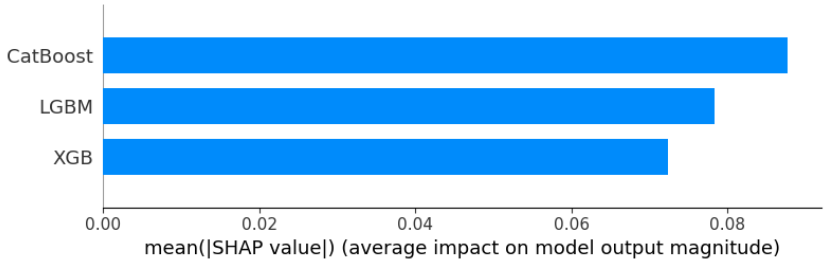


图 16 shap 变量重要性

b) 变量作用效果分析

绘制 LightGBM，CatBoost 的偏依赖图如图 17，可以看到随着两个模型的预测值的增加，Stacking 方法的预测值也会增加，相比 LightGBM，CatBoost 预测值的同等增加会引起最终预测值更大幅度的增加。

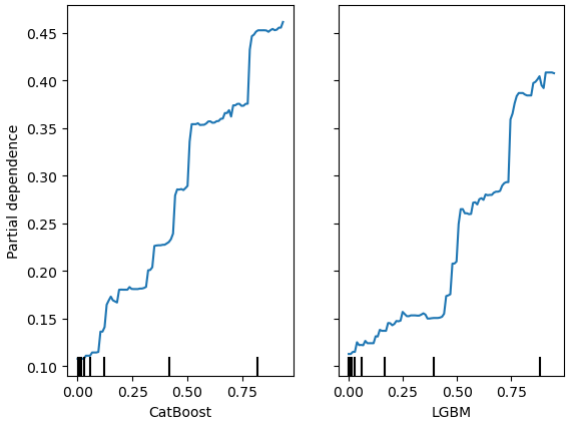


图 17 LightGBM，CatBoost 的偏依赖图

选取第一个观测样本，查看三个模型对其预测值的影响。在我们的预测最优 Stacking 方法中，base value 为预测值取 0 的概率，计算得其值为 0.176，即所有样本的平均预测为 1（患病）的水平为 0.176。如图 18，在该个体样本中，三个模型均使得预测值相对 base value 负向偏离。

图 18 第一个样本的 forceplot

查看三个模型对所有样本预测值的影响，如图 19，可以发现最优预测 Stacking 方法预测相对较为合理，相关性高的变量的预测值十分接近，尤其是被 Stacking 方法判断为不患病的样本。我们的方法对于不患病的情况预测较为极端，对于大部分样本，三个模型均倾向于使得预测值负向偏离 base value，且偏离值较一致。而对于少部分样本，三个模型的表现有一定分歧。对于被 Stacking 方法判断为患病的样本，三个模型的预测均倾向使得预测值正向偏离 base value，但偏离程度各有不同，其中，LightBGM 和 XGBoost 所导致的偏离值较为接近。

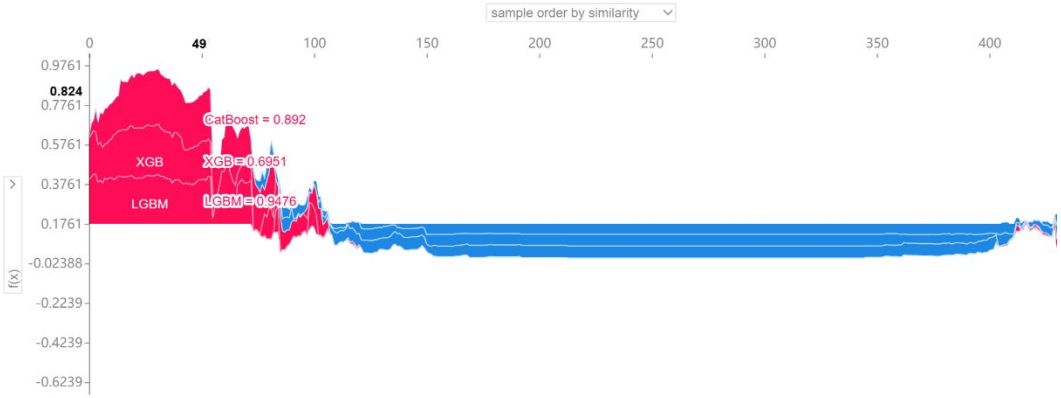


图 19 三个模型对所有样本预测值的影响

五、结论

(一) 模型评价

经过多次尝试多种数据预处理方法、单模型以及模型融合方法，我们的最优融合模型（基学习器为 LightGBM、CatBoost 和 XGBoost，元学习器为随机森林的 Stacking 模型）在验证集 1 上取得了 0.405 的 balanced log loss 值。这表明相较于单一模型，融合模型能够有效降低预测误差。在 LightGBM 模型效果解释部分，我们观察到匿名变量"DU"、"BQ"和"DA"对患病概率具有显著影响，并且进一步探究这些变量在不同样本上对患病概率的影响程度。这为人们根据易于测量的身体状态等指标来评估一个人是否有患病风险提供了有益信息，从而达到预防疾病的效果。

(二) 反思展望

1. 在最初划分训练集和测试集的时候，我们不慎造成了一些信息泄露，但是很快对此进行了改正，在今后的机器学习模型训练过程中也应多加注意这一问题。
2. 我们在探索过程中发现，尽管Lightgbm在单模型预测效果方面表现最优，但在所有的Stacking方法里Catboost都具有最高的变量重要性，我们后续还需要探究其中的原因。

3. 由于时间限制，我们在最终选择模型结果时没有进行多次训练集划分，只对一个验证机进行测试，并据此比较不同模型之间的表现，这种使用单一的验证集进行模型选择可能存在一定的风险，即这个结果可能不能反映真实的差异。在后续的工作中，我们应当在最外层的循中也循环划分训练集，在多组训练集和验证集上进行模型的训练和测试，以获得多次模拟的结果，并据此得到最终的最终的评估指标，从而使得最终的模型选择更加可靠和准确。
4. 我们在整个分析建模的过程中没有使用到greeks数据集，因为其内容与比赛介绍有所矛盾（greeks数据集没有标注患两种及以上病症的老年人）。后续若竞赛主办方给出合理解释和修改，我们会将greeks数据集的信息也纳入考虑，建立一个两步预测模型——先由健康特征数据信息预测患病具体情况（greeks数据集记录内容），再由此判断最终的Class类别。

参考文献

[1] COSCRATO V, de ALMEIDA INwidth=8,height=11,dpi=110CIO M H, IZBICKI R.The NN-stacking: feature weighted linear stacking through neural networks[J].Neurocomputing，2020，399：141-152.

[2] 方志. 基于 IGA-Optuna-LightGBM 的民航潜在旅客预测[J]. 国外电子测量技术. 2022,41(10)

[3] 李振华. 基于 Optuna-LSTM 的矿压预测方法研究[J]. 矿业研究与开发. 2023,43(03)

[4] 宋建, 王文龙. 基于 Stacking 集成学习的注塑件尺寸预测方法[J]. 华南理工大学学报(自然科学版). 2022,50(06)

[5] 张世文. 基于 Voting 策略的在线商品购买预测模型[D]. 湖南：湘潭大学, 2022

[6] 钱忆. 基于机器学习的 TPTV 用户报障预测算法研究[D]. 江苏：南京邮电大学,2018

[7] 朱益冬. 基于机器学习的信贷风控模型和算法研究[D]. 福建：厦门理工学院, 2023

附录

表 5 CatBoost 所选调优参数、调优范围及最优参数

参数名称	参数含义	范围	最优参数
iterations	可以构建的树的最大数量	(100, 1000)整数	689
learning_rate	学习率，用于减少梯度步长	(0.01, 0.3)小数	0.095
depth	树深	(3, 12)整数	3
l2_leaf_reg	代价函数的 L2 正则化项的系数，用于控制模型复杂度	(0.01, 1.0)小数	0.826
bagging_temperature	贝叶斯套袋控制强度	(0.0, 10.0)小数	7.443
random_strength	分数标准差乘数，用于防止过拟合	(0.0, 10.0)小数	3.098
border_count	特征二值化的数量，用于控制决策树叶节点数量	(1, 255)整数	212
scale_pos_weight	二值分类中第 1 类的权值，用于处理不平衡数据	sum(y_train1 ＝0)/ sum(y_train1 ＝1)	\

	集		
use_best_model	如果设置此参数，则定义结果模型中保存的树的数量	True	\
random_seed	随机数种子	48	\
logging_level	控制决策树叶节点数量的超参数	"Silent"	\
early_stopping_rounds	如果一个验证集的度量在 early_stopping_round 循环中没有提升, 将停止训练	200	\

表 6 随机森林所选调优参数、调优范围及最优参数

参数名称	参数含义	范围	最优参数
n_estimators	对原始数据集进行有放回抽样生成的子数据集个数	(100, 1000)整数	410
criterion	节点的划分标准	["gini", "entropy"]	“gini”
max_depth	决策树最大深度	(3, 12)整数	4
min_samples_split	节点可分的最小样本数	(2, 10)整数	9
min_samples_leaf	叶子节点含有的最少样本	(1, 5)整数	1
max_features	构建决策树最优模型时考虑的最大特征数	["sqrt", "log2"]	“sqrt”
class_weight	类型权重，用于处理不平衡数据集	"balanced"	\
random_state	随机数种子	48	\
early_stopping_rounds	如果一个验证集的度量在 early_stopping_round 循环中没有提升, 将停止训练	200	\

表 7 XGBoost 参数设置所选调优参数、调优范围及最优参数

参数名称	参数含义	范围	最优参数
objective	定义需要被最小化的损失函数	"binary:logistic"	\
n_estimators	集成的弱评估器的个数	(1000, 30000)整数	19808
learning_rate	学习率	(0.01, 0.3)小数	0.290
max_depth	树的最大深度	(3,12)整数	11
gamma	在树的叶节点上进行分支所需的最小损失减少量	(0.01, 0.7)小数	0.358
subsample	随机采样的比例，用于防止过拟合	(0.2, 1)间隔为 0.1 的小数	0.7
reg_alpha	L1 正则化参数，用于控制模型复杂度	(0.01, 0.7)小数	0.143
reg_lambda	L2 正则化参数，用于控制模型复杂度	(0.01, 0.7)小数	0.304

min_child_weight	最小样本权重和，用于防止过拟合	(0.1,10)小数	2.786
colsample_bytree	用来控制每棵随机采样的列数的占比	(0.1,0.9)小数	0.687
max_delta_step		(0,10)整数	6
alpha		(0.01,0.7)小数	0.432
early_stopping_rounds	如果一个验证集的度量在 early_stopping_round 循环中没有提升，将停止训练	200	\
random_state	随机数种子	48	\

表 8 SVC 参数设置所选调优参数、调优范围及最优参数

参数名称	参数含义	范围	最优参数
C	惩罚系数	(1e-3,1e3)对数域	0.077
kernel	核函数类型	"linear","rbf","sigmoid",”poly”	“rbf”
gamma	核函数系数，只对”rbf”,”poly”,”sigmoid”起作用	"scale","auto"	“scale”
degree	多项式核的阶数	(1,5)整数	1
coef0	核函数的常数项，只对'poly','sigmoid'有用	0	\
class_weight	类型权重，用于处理不平衡数据集	“balanced”	\
random_state	随机数种子	48	\

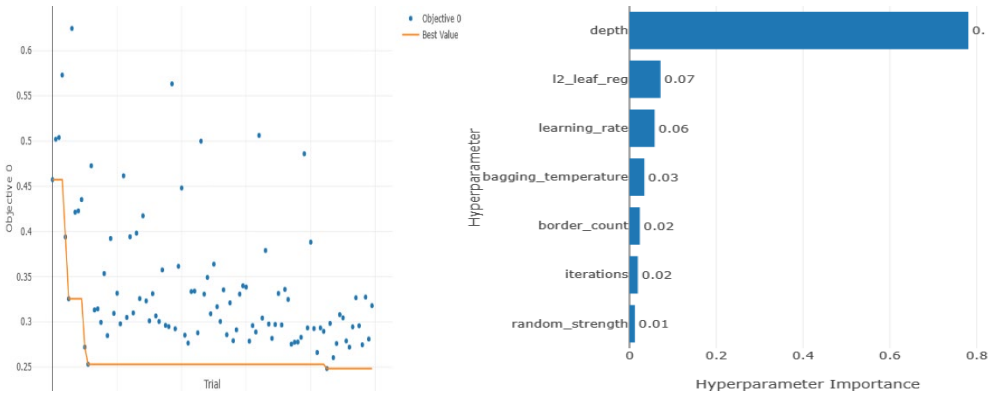


图 10 基于 Optuna 优化框架下的 CatBoost 调参过程

（左图为目标函数值随迭代次数的变化情况，右图为参数重要性）

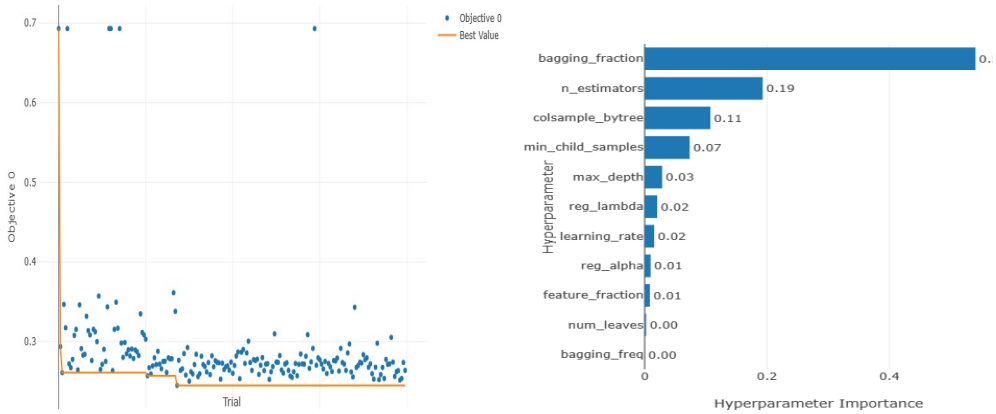


图 11 基于 Optuna 优化框架下的随机森林调参过程

（左图为目标函数值随迭代次数的变化情况，右图为参数重要性）

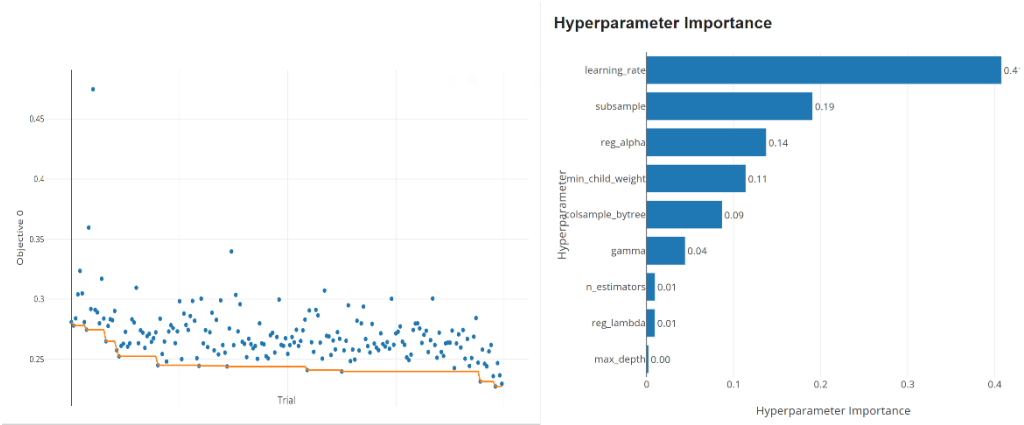


图 12 基于 Optuna 优化框架下的 XGBoost 调参过程

（左图为目标函数值随迭代次数的变化情况，右图为参数重要性）

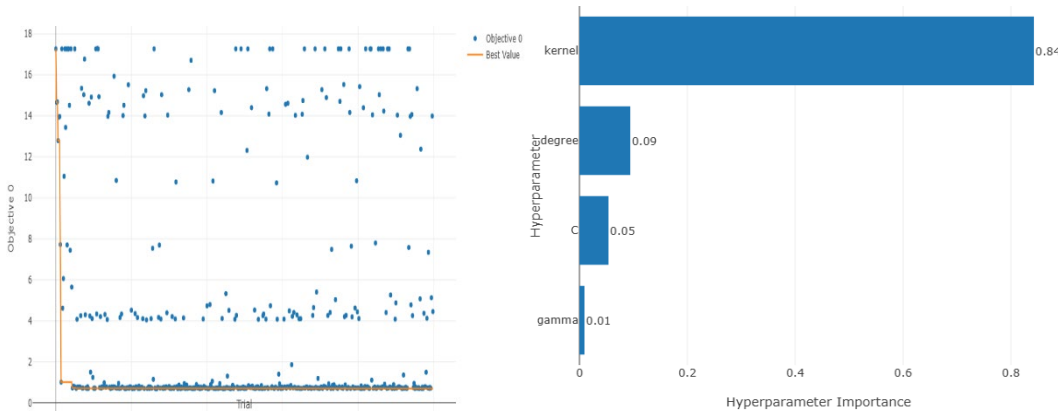


图 13 基于 Optuna 优化框架下的 SVC 调参过程

（左图为目标函数值随迭代次数的变化情况，右图为参数重要性）