模型引入

GBDT 的计算方法说明

梯度增强决策树是一种由多颗决策树组成的迭代式决策树，其最终的决策效果是用加的方法得到的。GBDT 在每个迭代过程中都会增加一个新的决策树，从而使其更加精确。利用正态分配和相加模式实现了最优的学习。其基本过程如下：一是对一个仅有根结点的树状结构进行初始化；然后构造 M 个基本学习程序，通过计算损失函数来估计剩余量；建立一个返回树形 CART 来对剩余进行拟合；采用树叶结点进行拟合，以最小化损失为目标；最终，升级学习程序

GBDT 的运算过程：

（1)初始化基础训练程序 $f_0(x)$,为式

$$f_0(x) = \arg\min_c \sum_{i=1}^{N} L(y_i, c)$$

上式中 $L(y_i,c)$为损失函数，用于计算真实值与预测值之间的误差，argmin 为确定损失函数值最小时 c 取值的函数。

(2)在此基础上，建立了一个损耗方程，用以求出实际值和预报值的偏差，以求出最小损耗函的最小值。

$$f_0(x) = \arg\min_c \sum_{i=1}^{N} L(y_i, c)$$

(3)通过构造一组 CART 的线性关系，采用斜率上升技术对残差进行拟合

GBDT 则把损失的负向斜率当作残差估算记为 r_{mi}，并给出一个公式

$$r_{\mathrm{mi}} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right] f(x) = f_{m-1}(x)$$

(4)在确定剩余估算后，采用 CART 回归树法进行拟合法，获得了各叶片结点区 R；最优拟合的相应损耗函数是一个方程

$$C_{mj} = \arg\min_c \sum_{x_i \in R_{mj}} L(y_i, f_{m-1}(x_i) + c)$$

(5)将学习机升级为一个公式

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} c_{mj} I\left(x \in R_{mj}\right)$$

在这里，I 是一个学习率

$$L(y, f(x)) = \sum_{y \geq f(x)} \theta |y - f(x)| + \sum_{y < f(x)} (1 - \theta) |y - f(x)|$$

在公式中，θ为分位数

上面是一个板块，下面是另一个板块

模型的建立与计算

分析的步骤

1。利用训练集数据建立梯度提升树（GBDT）分类模型。

2。通过建立的梯度提升树（GBDT）计算特征重要度。

3。将梯度提升树分类模型（GBDT）运用于训练，测试数据中，获得模型分类评价结果。

4。由于梯度提升树（GBDT）的随机性，每一次操作都会有不同的效果，如果将当前训练模型进行保存，后续可将数据直接上传到当前训练模型中，用于计算分类。

5。注：梯度提升树（GBDT）不能象传统模型那样获得确定方程，模型的评估一般采用测试数据的分类效果

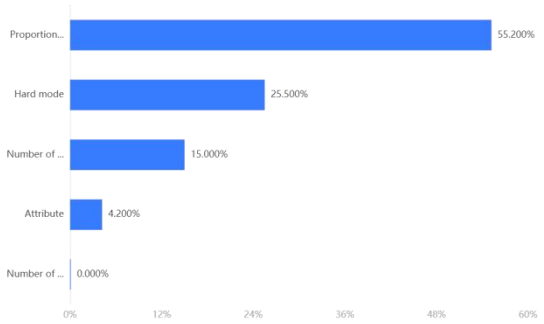
模型参数

参数名	参数值
训练用时	0.672s
数据切分	0.7
数据洗牌	否
交叉验证	否
损失函数	deviance
节点分裂评价准则	friedman_mse
基学习器数量	100
学习率	0.1
无放回采样比例	1
划分时考虑的最大特征比例	None
内部节点分裂的最小样本数	2
叶子节点的最小样本数	1
叶子节点中样本的最小权重	0
树的最大深度	10
叶子节点的最大数量	50
节点划分不纯度的阈值	0

图表说明：

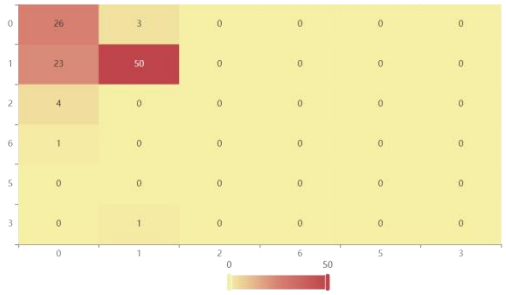
上表展示了模型各项参数配置以及模型训练时长。

特征重要性



图表说明：

上柱形图或表格展示了各特征（自变量）的重要性比例。



混淆矩阵热力图

图表说明：

上表以热力图的形式展示了混淆矩阵。

模型评估结果

	MSE	RMSE	MAE	MAPE	R^2
训练集	0.98	0.002	0.001	76.534	1
交叉验证集	0.812	0.773	0.509	588303.205	-0.721
测试集	1.22	1.105	0.82	67422.005	-1.234

图表说明：

交叉验证集显示在上表上、在训练集与测试集上预测评价指标，采用量化指标对 GBDT 预测效果进行测度。在此基础上，利用遗传算法对该方法进行优化处理，使得算法能快速有效地收敛到最优解。其中超参数可由交叉验证集评价指标连续调节，为了获得可靠，稳定的模型。

●MSE（均方误差）：预测值和实际值差值平方的期望值。该指标可以作为衡量预测结果准确性和可信度的重要标准之一。取值越小模型的准确度越高。

●RMSE（均方根误差）：为 MSE 的平方根，取值越小，模型准确度越高。

●MAE（平均绝对误差）：绝对误差平均值，能够较好地反映预测值误差真实状况。在给定精度要求下，取不同数值时相对偏差值与相对标准偏差之差即为预测结果的准确程度。取值越小模型的准确度越高。

●MAPE（平均绝对百分比误差）：是 MAE 的变形，它是一个百分比值。取值越小模型的准确度越高。

●R2：将预测值跟只使用均值的情况下相比，结果越靠近 1 模型准确度越高。

对于 1—7 次我们都进行这样的一次训练，我们分别得到了以上图表，分别输入数据，可以得到想要的预测值。输入 2023 年 3 月 1 日的单词 EERIE，我们得到以下表格：

为了保证模型选择的准确性，我们又选取了 2 个模型来进行拟合优度对比分析，结果见表。

R-squared	DTA	ET	GBDT
R^2	0.8546	0.8843	0.9232

拟合优度 R^2 最大值为 1。R^2 的值越接近于 1,说明当前回归方程对预测值的拟合度越好，观察表 GBDT 模型的拟合度最高，使用此模型有较高的可靠性。

1try	2tries	3tries	4tries	5tries	6tries	7tries	重复字母个数
0.9867	5.8440	10.3849	28.5336	34.9876	14.2832	4.5484	3

对于以上数据求和是 99.5675，非常接近 100。并且集训表现均在 90%以上，精度很高，因此我们创立的模型切合度很好。

此外，考虑到时间问题，2023 年 3 月 1 日的最终预测报告

参考文献

[1] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com.

[2] 周志华. 机器学习[M]. 清华大学出版社, 2016.