**数据集说明**

整体而言，我们的数据集由2011年和2015年两个时间段构成。其中，2011年含10027个样本，2015年含9668个样本。这两个数据集中均包含有匿名ID以及householdID来关联到每一个数据样本提供者。同时，除了ID之外，还包含有19个血液生物指标，14个慢性病指标，20个功能障碍指标，8个认知抑郁指标以及我们的目标变量年龄。

**分析流程**

1. 数据清洗

首先，针对年龄特征过滤出在40至85岁（含40和85）年龄段的样本。第二，对数据集做去重去空操作。第三，针对每个特征列计算数据缺失率，以此来作为衡量是否需要舍去对应列的一个标准（通常会舍去具有较高数据缺失率的特征列）。最后，检查各特征列的数据类型，确保数值类型特征与类别类型特征准确定义。

1. 特征工程

* 缺失值填充（通常用均值、中位数、众数等）
* 区分出数值类型，类别类型，序数类型数据后，对类别变量采用独热编码，对序数变量采用序数转换。
* 标准化（Min-Max scalar 或者 standard scalar）
* 去除低方差特征
* 去除多重共线性

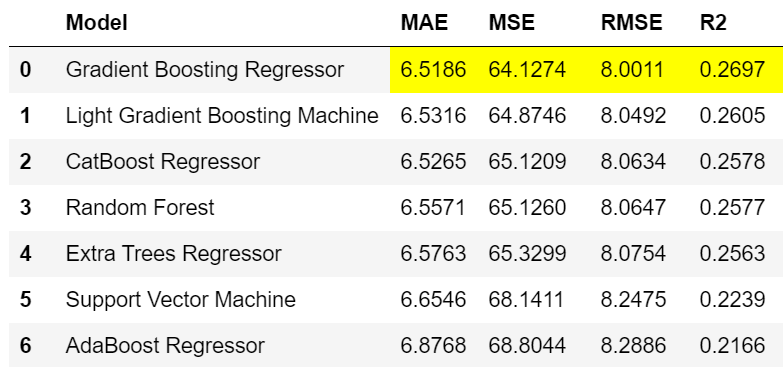
1. 模型比较

为了证明非血液指标特征对预测年龄有重要作用，我们首先只用血液指标特征来训练模型，之后我们在血液指标变量的基础上加入其他非血液指标变量来重新训练模型，通过比较两次训练出来的模型的R-square和MAE(Mean Absolute Error)值来说明非血液指标变量对预测年龄的重要性。

1. 结果

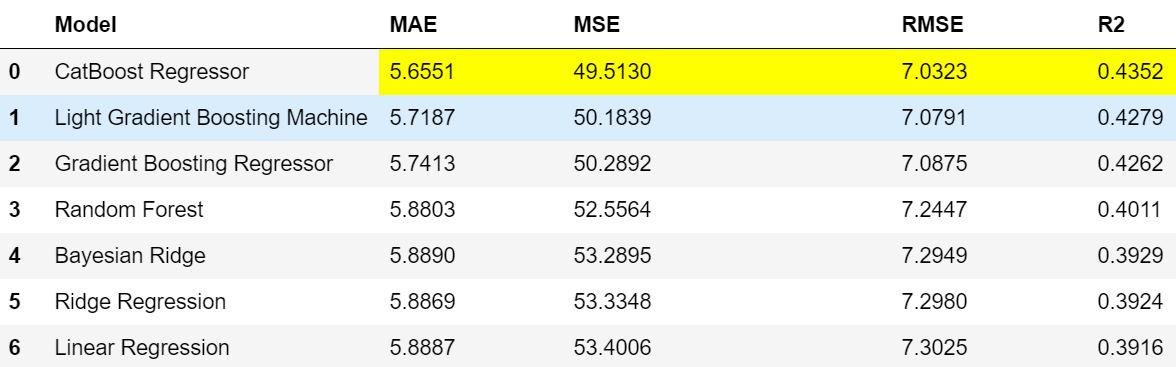
通过对现有的regressor类模型分别做十折交叉验证并训练，将得到的模型表现从高到底排序，依次为：

* 仅血液指标



表现最好的模型是Gradient Boosting Regressor, R-Square = 0.2697, MAE=6.5186

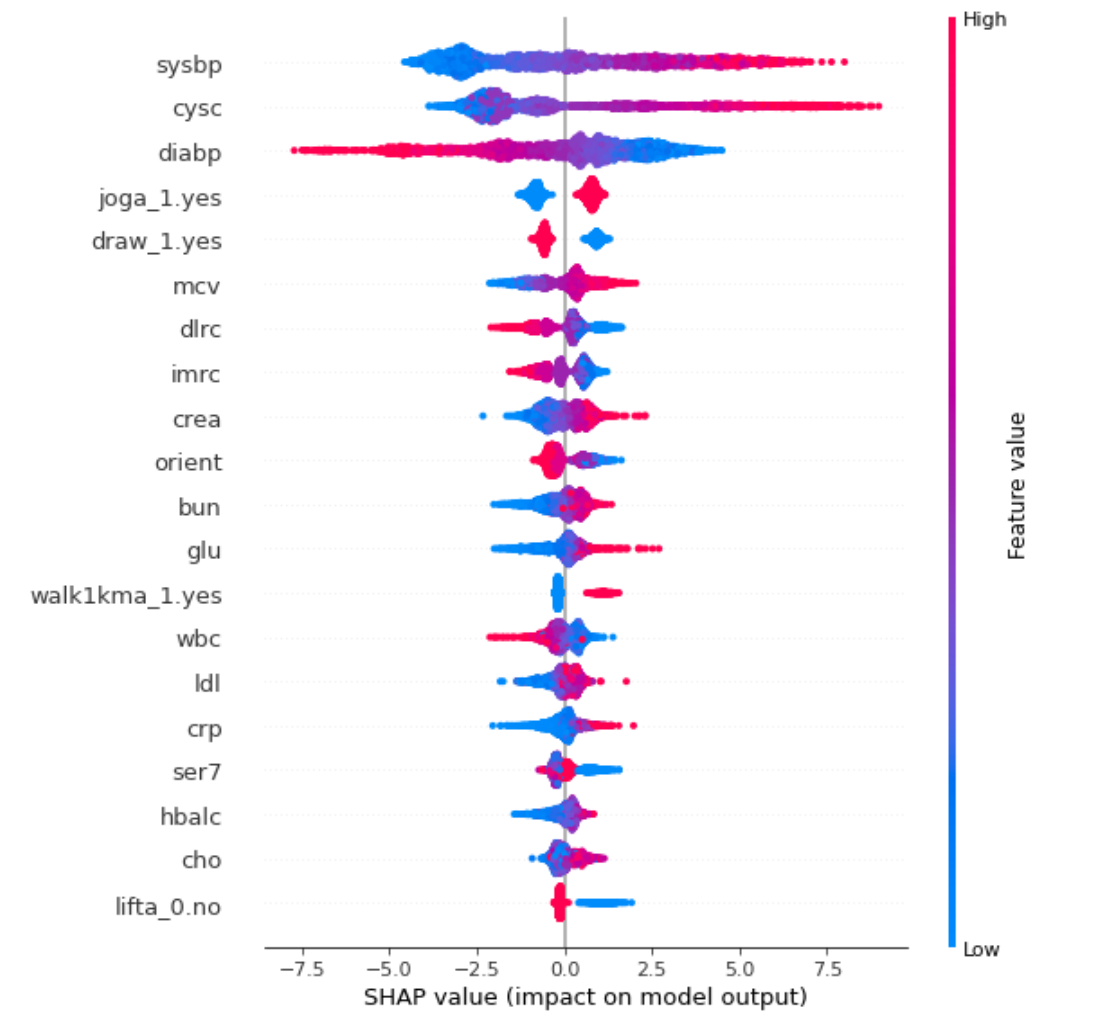
* 在血液指标的基础上加入其他指标



表现最好的模型是CatBoost Regressor, R-Square = 0.4352,MAE=5.6551.

1. 特征对年龄预测的影响

通过SHAP Value分析，可以看到对年龄预测有较大影响的几个特征变量。SHAP Value是一种基于博弈论的被用来解释机器学习模型中输入变量与预测值之间的关系。



Y轴方向为feature value, 颜色从蓝到红，代表数值从低到高，X轴方向为SHAP值，同时特征变量的重要性从上到下依次排列。

重要性最高的5个变量依次是：sysbp(systolic blood pressure), cysc(cystatin C), diabp(diastolic blood pressure), joga(some diff-jog 1km),draw(cognition able to draw assign picture).较高的systolic blood pressure或者cystatin C 会对预测年龄有正向作用，而较高的diatolic blood pressure对预测年龄有负向作用。而joga和draw出现在其中则说明，非血液指标对于预测年龄同样具有重要意义。

**结论**

通过比较加入非血液特征变量前后的机器学习模型表现，我们发现，R-square从0.27 上升到0.44, MAE从6.52年降低至5.66年，这表明非血液指标能够在一定程度上提高机器学习模型预测年龄的能力。同时，SHAP Value的方法也从侧面反映出来非血液指标在年龄评估预测方面的重要性。