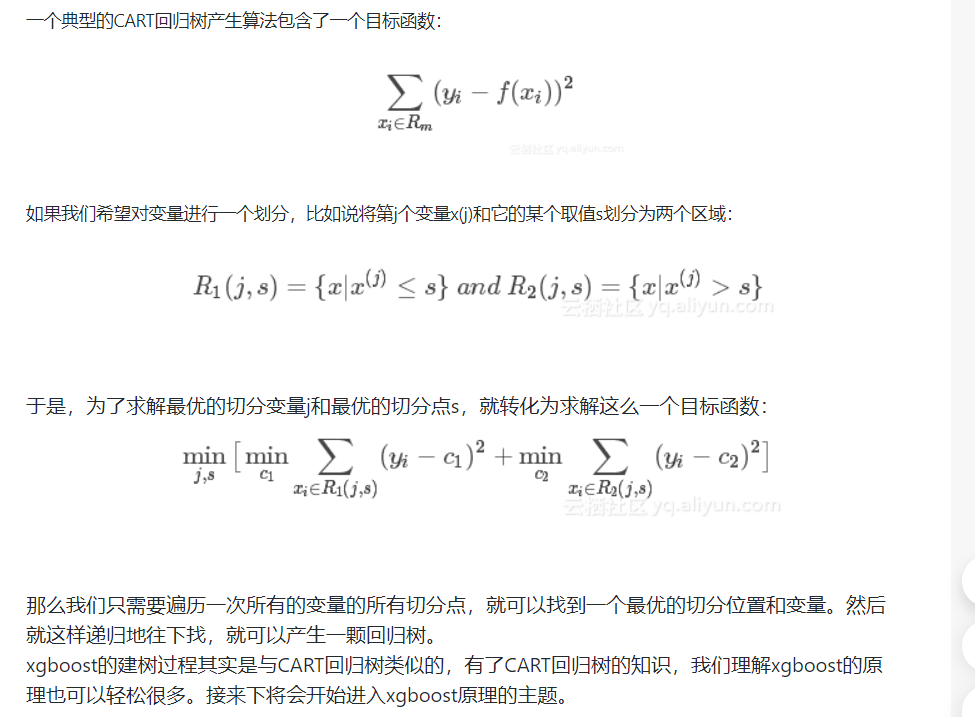
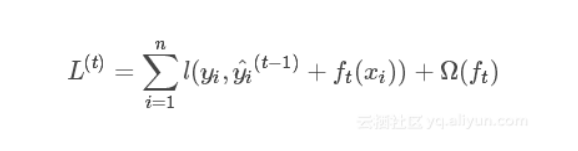
XGBOOST：

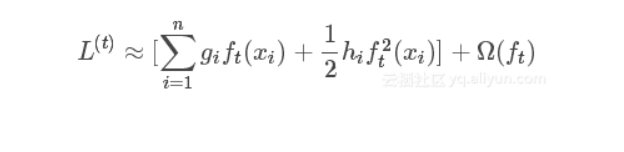
由于一共有11个特征，通过皮尔森相关系数他们并没有很大的相关性（放皮尔森相关系数热力图），对其降维的可操作性不大，各个自变量与因变量的相关性复杂，不能使用多元线性回归模型来进行拟合寻找关系。因此我们选择了基于决策树的回归模型。

（决策树的数学原理如下图）：

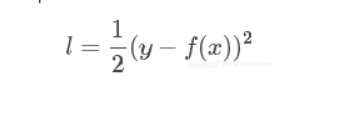
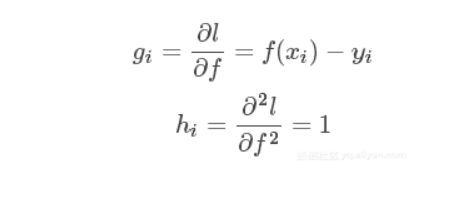
但是仅仅使用一棵决策树通常并不能使拟合的效果达到很好，所以我们想利用计算机迭代计算的能力来使用多棵树并行操作对数据的拟合。相比于平行使用决策树，并行操作能够使得我们在一棵树决策完之后对这棵树的结果进行评价，通过贪婪算法（局部最优解）降低loss值，使得每次迭代的同时都能使loss减小，这样后续树就能减少前一棵树的错误。因此我们定义了一个针对每一棵树的评价函数：



为了更快地去优化这个函数，我们使用梯度下降法去优化损失函数。我们可以在ft=0处二阶泰勒展开，并且目标函数不受上一次的loss的影响，于是我们删除loss项目得到下图：

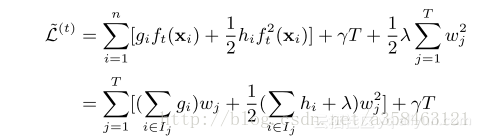


其中gi表示L对f的一阶偏导数，hi表示l对f的二阶偏导数，因为是回归任务，我们将损失函数定义为MSE（mean square error）（公式如下）：

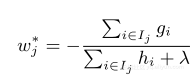


所以

由于每个ft(xi)都对应着一个叶子结点wi，于是我们可以用wi来代替一个ft，所以我们将该目标函数改写一下可以得到：



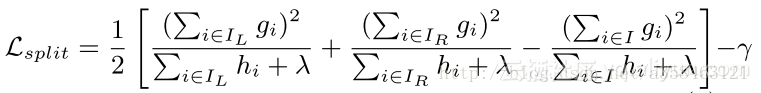
我们利用导数为0的极值性质对wj求偏导，得到基于该目标函数的最优权重：



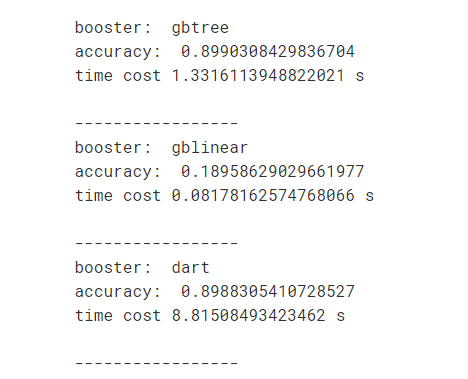
我们将最优权重代入回原函数中，就得到我们对决策树ft性能的最终评分函数：

image

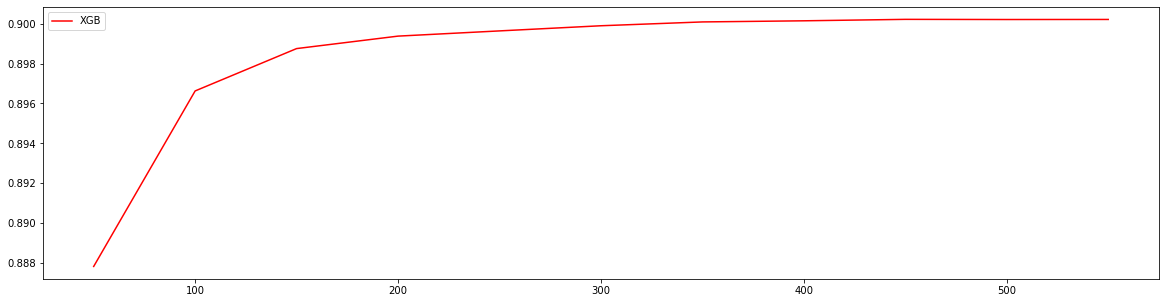
算法的基本步骤应该就是遍历所有特征的所有分割方法，选择损失最小的，得到两个叶子，然后继续遍历。遍历的时候可以并行的执行。总公式如下图：



我们将这个基于许多决策树而形成的集成学习方法称为Boosting，将每棵树称为基学习器（Base Learner）。在本次实验中，我们首先挑选了决策树，回归，和dart树来作为基学习器分别对比准确率和运行时间，我们发现决策树和dart树的准确率相符并且运行时间决策树更加快，所以我们选择决策树来作为我们的基学习器。

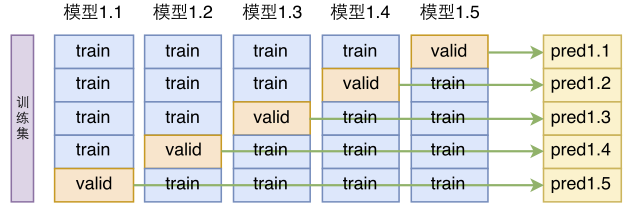


并且在使用Boosting来进行拟合的时候，基学习器数量的选择至关重要，它关系到我们Boosting模型的复杂度，基学习器的数量太少容易欠拟合（underfitting）（学习不到特征，回归结果差），基学习器过多，模型会过于复杂。



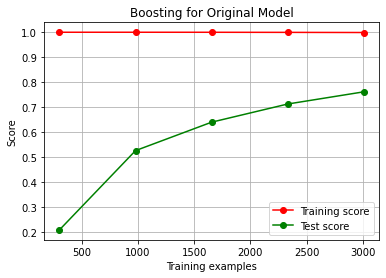
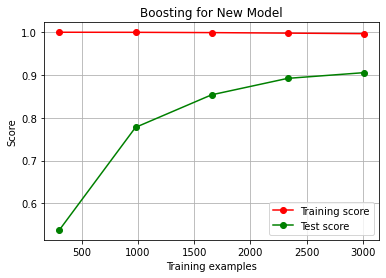
在拟合的时候，随着基学习器的增多，拟合的准确度在上升，在达到400左右的时候已经趋于收敛的状态，再多的基学习器也并不能帮助我们更好的决策，所以我们将基学习器的数量定在400左右。

为了能够评估模型对数据的拟合程度，我们将数据分为训练集（training data）和测试集（testing data），其中测试集不参与训练，用于最终模型的评估。并且我们在决定boosting模型参数的时候（包括基学习器数量和基学习器种类），是通过在训练集中划分出一部分训练数据作为验证数据(Validation data)来进行调参，最后再通过测试集进行评估，这样可以相对客观地评估模型对于训练集之外数据的匹配程度。我们使用的是K则交叉验证（K-Fold cross validation）。



它将原始数据分成K组，将每个子集数据分别做一次验证集，其余的K-1组子集数据作为训练集，这样会得到K个模型。这K个模型分别在验证集中评估结果，最后的误差MSE(Mean Squared Error)加和平均得到交叉验证误差。其利用了有限的数据，并且评估结果能够尽可能接近模型在测试集上的表现。

最终我们在使用10个自变量对原模型和新模型的评分进行回归拟合，对原模型的拟合效果达到了76%左右，平均MSE在0.09左右；在对新模型的拟合效果达到了90%左右，平均MSE为0.04。

并且原模型和新模型的权重比为（高亮一下以下指标）：

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Weight | Total Energy | Fertilizer | Government\_Management | Law&Political Stability | Unemployment\_rate | Yield | Gini\_Index | Male/Female | Rural/Urban | Developed |
| Original | 17% | 6% | 13% | 14% | 6% | 11% | 9% | 9% | 10% | 6% |
| New | 26% | 6% | 9% | 2% | 7% | 5% | 7% | 12% | 10% | 17% |

我们可以发现无论是原模型还是强调equality和sustainability的新模型，农业能源使用这个指标都是非常重要的衡量标准，并且在评价equity and sustainability其重要程度更大，而其他指标对于模型的贡献较为平均。关注efficiency and profitability的原模型更加注重政治方面的发展以及农业产出，他们对于维系一个国家政治发展的稳定和农业中经济发展的稳定有着较高的要求，并且随着政府管理能力的增强，对政治的重视程度提高，以及产量越大，其原模型的得分越高。

将食物模型改为equity and sustainability orient的话需要更加重视Total Energy use，所以对于可持续性发展来看，我们需要减少能源排放的使用。由于国家的发达程度也是equity and sustainability orient的食物模型的一个重要衡量标准，其意义在于随着一个国家发达程度的提高，其可持续性发展水平是更高的。结合实际来看想要实现可持续性发展的均衡，需要通过发达国家来带动发展中国家共同发展，加强发展合作充分调集资源，向发展中国家提供资源帮助。联合国于 2015 年正式提出可SDGs，要充分认识到发展必须建立在平衡社会、经济和环境的可持续性之上。并且也间接说明了目前发达国家和发展中国家在equity and sustainability orient的食物系统中有着不同的表现。进而我们通过发达国家和发展中国家进行区分，分别训练计算权重得到分类之后的权重占比如下表所示：

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Weight | Total Energy | Fertilizer | Government\_Management | Law&Political Stability | Unemployment\_rate | Yield | Gini\_Index | Male/Female | Rural/Urban |
| Original | Developed | 13% | 9% | 6% | 17% | 4% | 5% | 20% | 19% | 8% |
| developing | 19% | 6% | 10% | 11% | 8% | 9% | 8% | 10% | 18% |
| New | Developed | 33% | 3% | 13% | 3% | 1% | 6% | 2% | 27% | 12% |
| developing | 32% | 5% | 11% | 3% | 8% | 7% | 8% | 10% | 16% |

无论是重视equity and sustainability 还是efficiency and profitability，男女比例对发达国家的影响都是比发展中国家大的（蓝色字部分），并且农村城镇人口比和失业率对发展中国家的影响是比较大的。

在efficiency and profitability的食物系统中，政治因素，基尼指数对一个国家efficiency and profitability食物系统的衡量起着重要作用，这也说明了发达国家实现efficiency and profitability的发展是将重心放在了政策和国民收入稳定性上面。其中基尼指数较低代表贫富差距越小，并且贫富差距越小的国家主要提升了食品系统的效率，进而使得国家在原模型中得分更高。而发达国家也主要通过政策对国家稳定性的影响力来促进profitability的发展。而在efficiency and profitability针对发展中国家的评价中，Total Energy use对评价efficiency and profitability有着重要的影响，这说明与发达国家不同，发展中国家想要实现效率的提升和利润的改变，需要更重视能源的使用。

所以要实现从efficiency and profitability orient向equity and sustainability orient进行转化，需要加大对Total Energy use的重视，并且实现发展中国家向发达国家进行转化，并且需要兼顾适量产出和政策导向。对于发达国家来说，需要带动发展中国家的发展，需要进行资源的调度来缩小世界上国家之间的贫富差距，并且在发展的同时需要兼顾环境保护和可持续能源的使用，需要减少能源的排放。对于发达国家来说，由于工业化水平的发展，短时间内减少能源的排放不太现实了，这就需要发达国家为了世界环境的良好运营改变原本利润最大化的目标，牺牲一小部分利益来促使食品系统平等和可持续性的发展。发展中国家在转化的同时将会获得部分发达国家的帮助，但是对内需要调整人口水平，兼顾环境保护的同时调整农村城市人口结构，让农村人口向城市进行转化，同时也要通过政策福利来减少国家的失业率。对于发展中国家来说实现食品系统健康的发展需要从经济以及社会的角度来进行改变，提升国家的工业化水平，提高国家的粮食产量以及经济发展等。但是在工业化的同时也会对生态环境造成一定的影响，所以需要balance生态和工业化的发展来达到一个良好的食品系统。

在我们efficiency and profitability orient模型中，由于我们认为一个能够实现自给自足并且能够保持一定生产力和维系较低的食品通货膨胀的国家为较为良好的食品系统，在20年的发展之中，从00年-10年发达国家的生产力相对来说较强，食品评价系统得分较高，但是在近十年来，发达国家的通胀水平以及工业化的转变使得发达国家不将重心放在了食品系统上面，更加重视经济、科技的发展，偏向从部分较为发达的发展中国家以及一些粮食出口国（例如巴西等）来进口粮食，食品系统的得分就会慢慢降低，而部分发展中国家由于近十年逐步发展，成为了能够自己自足甚至粮食出口的国家，所以在近十年的得分当中，发展中国家的得分较高。

在我们equity and sustainability orient中的模型当中，我们发现大部分发展中国家，包括一些非洲的国家表现的较为良好，主要是这些国家工业化水平不高，在可持续方面有着良好的评价，并且一些欧洲的岛屿国家由于可再生资源丰富，对环境保护的工作也做的非常良好，同样得分很高