**Python**基于音频数字信息智能诊断帕金森疾病

1. 项目背景

任务描述：

帕金森疾病是一种常见的老年神经系统退行性疾病，但其确切病因目前仍不清楚，而传 统的疾病监测需要专业人员进行复杂的操作，并且监测周期长，所以对于帕金森疾病的检查 和诊断仍无法实现快速检测的需求。近年，在医学研究上已证明基于语音数据实现帕金森疾 病诊断是一种有效的方法，但诊断系统的算法不优，不能满足实际疾病诊断需求，而且系统 的分类准确率域稳定性不佳。 构建一个有效的帕金森疾病诊断系统，可实现对患者进行远程诊断和病情监控，也可作 为医生治疗或诊断的辅助工具，且能够作为帕金森疾病预防的自我检测手段。

任务分析：

利用人工智能技术开筛选多种与疾病相关的语音特征，运用人工智能算法 构建帕金森疾病诊断模型，使用评估方法对比多种算法的诊断效果，选取最优模型。

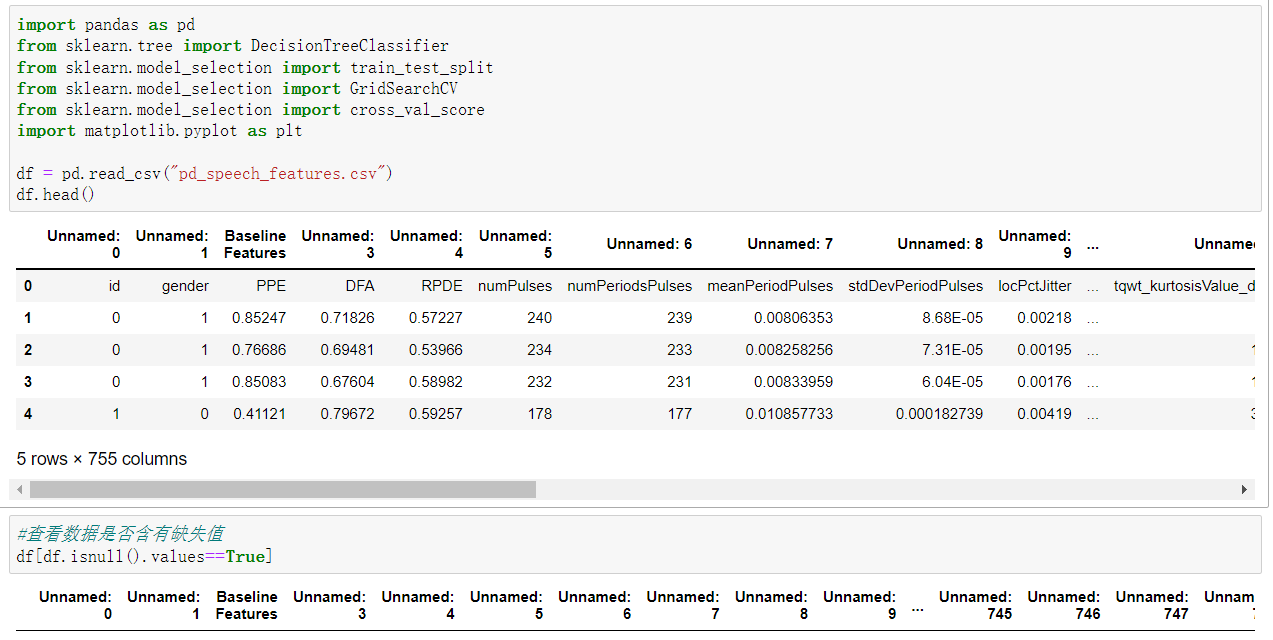
2. 数据读取与特征探索

任务描述：

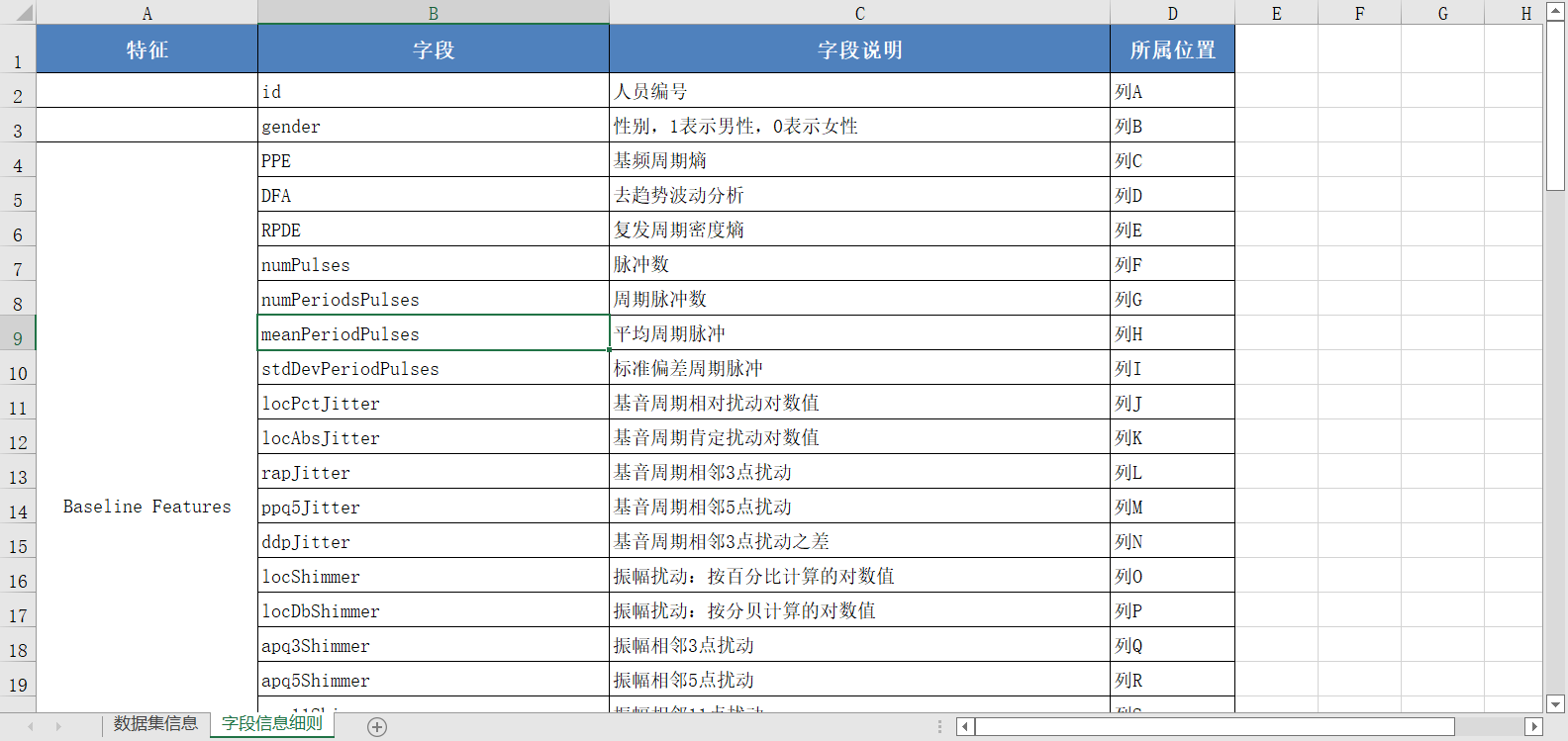
数据从第二行开始才是数据具体的字段名称，数据读取 需注意该情况，由于数据含有特征数量较多，需对数据进行详细的探索，掌握所提供音频数 字信息的情况。

2.1 读取数据并挑选有用信息

通过前五行的数据了解到该表的一定信息，该模型共有七百多条特征，且无缺失值



通过字段说明的excel了解了各列所代表的意义



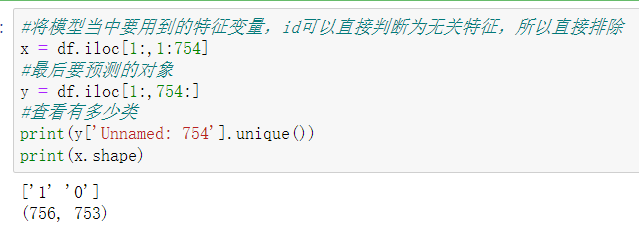
3数据预处理与特征筛选

任务描述：

对数据进行简单的预处理，筛选与帕金森疾病相关的 特征，减少模型训练所需的特征数量。

3.1对数据进行划分

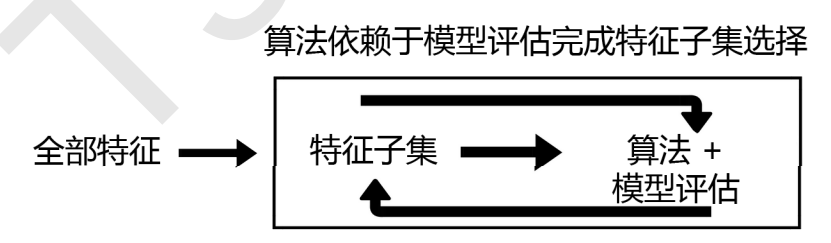
将特征变量放入x中，将分类的结果放入y中，并观察y共有多少类型以及x的大小



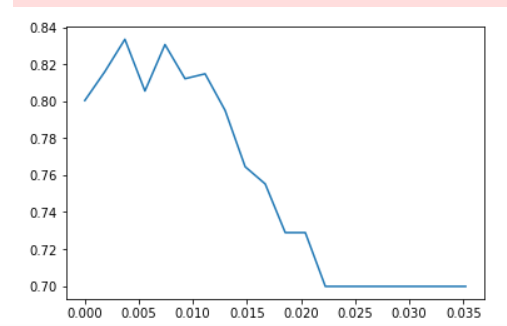
3.2特征筛选

在无法依赖对业务的理解来选择特征的情况下，本项目通过嵌入法来对数据的特征进行筛选。

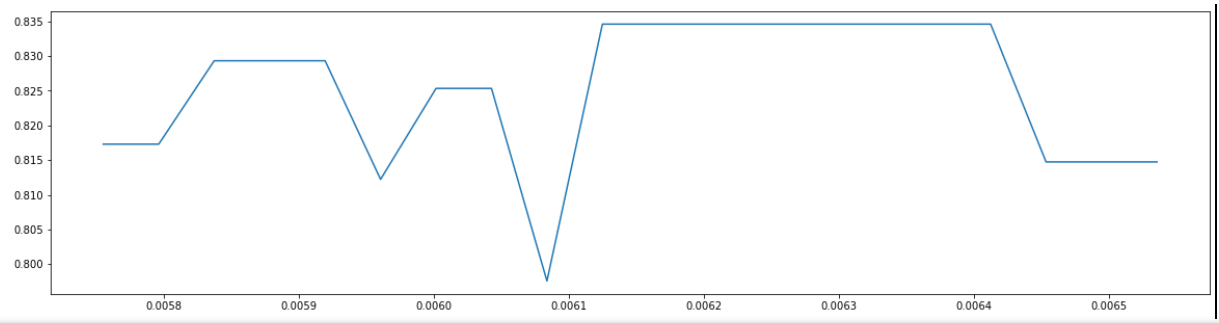
嵌入法是一种让算法自己决定使用哪些特征的方法，即特征选择和算法训练同时进行。在使用嵌入法时，我们先使用某些机器学习的算法和模型进行训练，得到各个特征的权值系数，根据权值系数从大到小选择特征。这些权值系 数往往代表了特征对于模型的某种贡献或某种重要性，在树的集成模型中的feature\_importances\_属性，可以列出各个特征对树的建立的贡献，我们就可以基于这种贡献的评估，找出对模型建立最有用的特征。



由于嵌入法的使用的权值系数没有指定的范围可找,可以说,权值系数为0的特征对模型丝毫没有作用，但当大量特征都对模型有贡献且贡献不一时，我们就很难去界定一个有效的临界值。这种情况下，模型权值系数就是我们的超参数,所以需要画出学习曲线来判断这个超参数的最佳值究竟应该是多少。



通过学习曲线可以清楚的看出在0-0.005之间模型的效果最好，因此我们就着这一区间进行缩小我们的范围



最后确定最佳权值系数为0.00612515，使用其构建的模型准确度达到目前最大值

0.8345964912280701

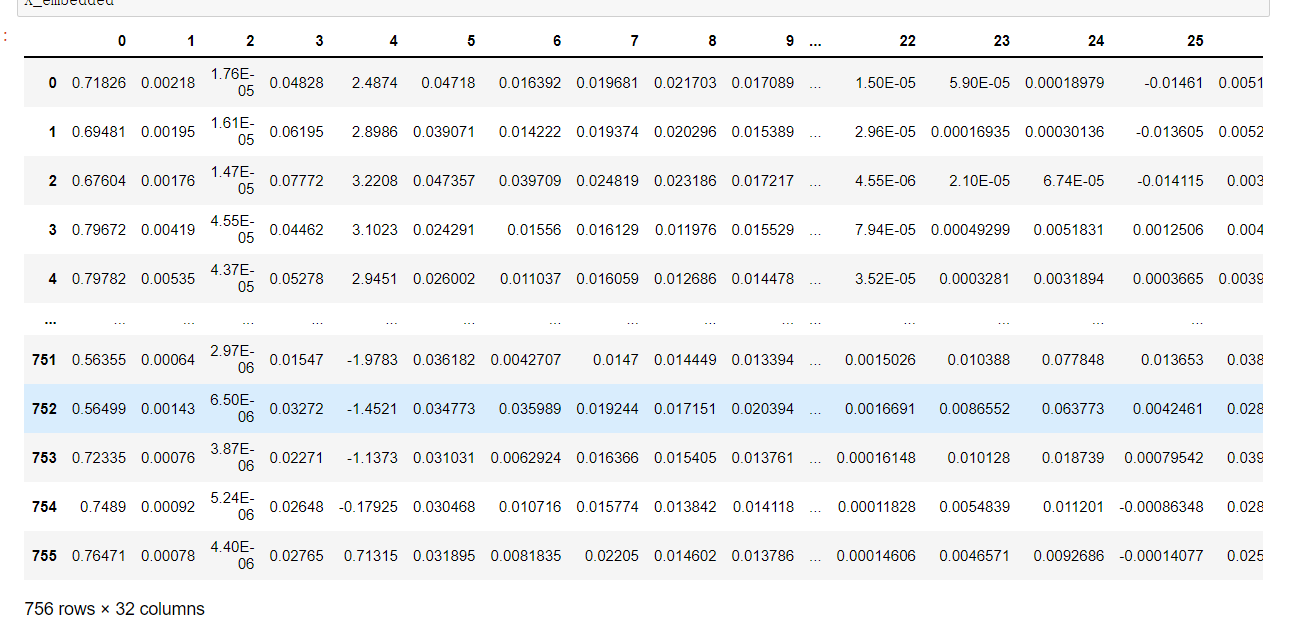
4特征可视化分析与标椎化

任务描述：

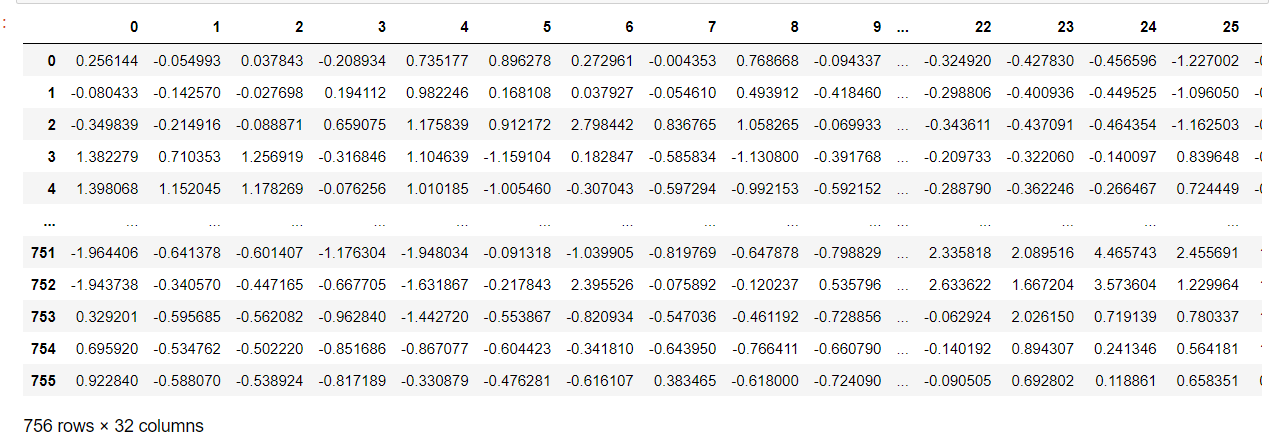
对重要的特征进行可视化分析，了解数据分布情况， 根据可视化结果实现数据标准化。

4.1可视化与标椎化

经过前面特征筛选后七百多列特征最终只剩下32列重要特征，通过DataFrame将其可视化，其结果如下



进而对其进行标椎化



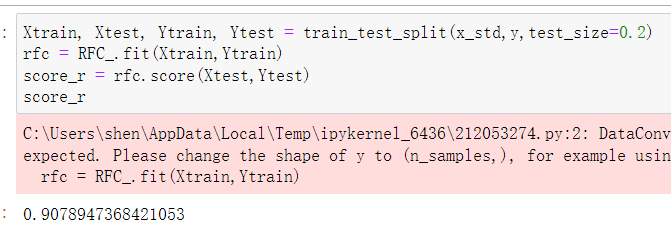
5模型构建与训练

任务描述：

选取分类算法搭建帕金森诊断模型，使用处理后的数据 进行模型训练。

5.1构建随机森林模型

将模型进行8:2分为训练集与测试集。模型初步的准确度达到90%左右，算是蛮好的开始。



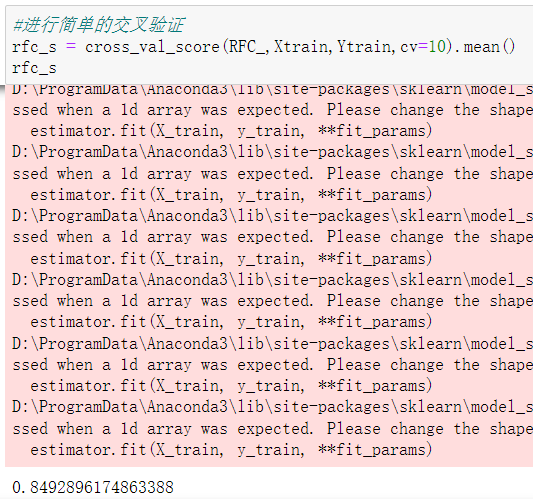
6模型评估与优化

任务描述：

选取分类模型的评估算法判断模型的优劣情况，并进行 模型优化，最终确定帕金森诊断效果较优的模型，该过程需明确给出各模型的评估结果。

6.1交叉验证评估模型

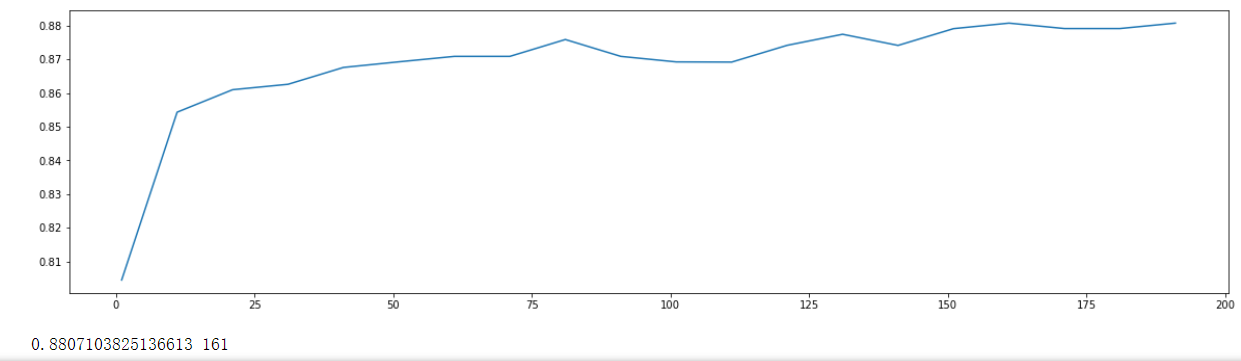
交叉验证是用来观察模型的稳定性的一种方法，我们将数据划分为n份，依次使用其中一份作为测试集，其他n-1份 作为训练集，多次计算模型的精确性来评估模型的平均准确程度。训练集和测试集的划分会干扰模型的结果，因此 用交叉验证n次的结果求出的平均值，是对模型效果的一个更好的度量。



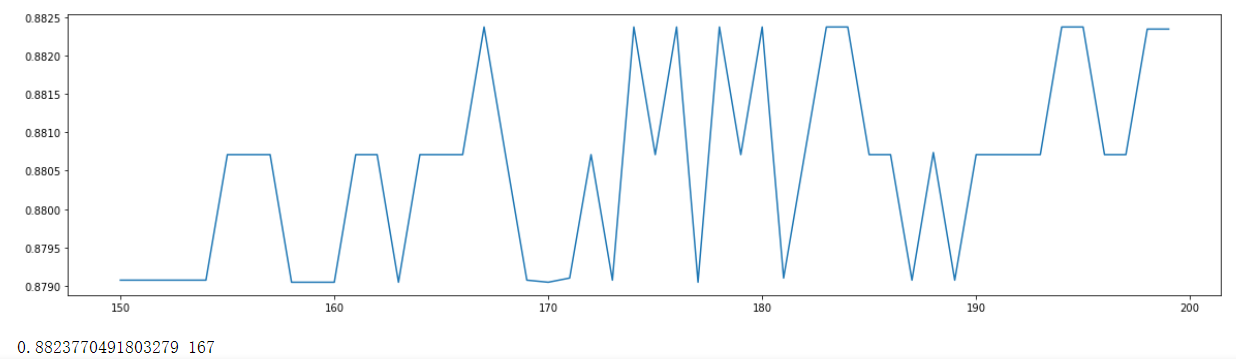
在对训练集进行交叉验证后得出的结果反而没有测试集高（不清楚原因）

6.2调整随机森林中树的数量

绘制每隔10个取一个值的学习曲线观察n\_estimators的范围，当n\_estimators=161时，模型的准确度在0-200区间达到最大



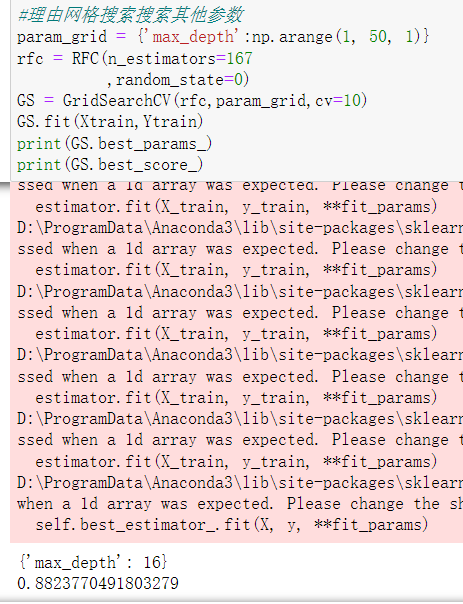
进一次缩小范围



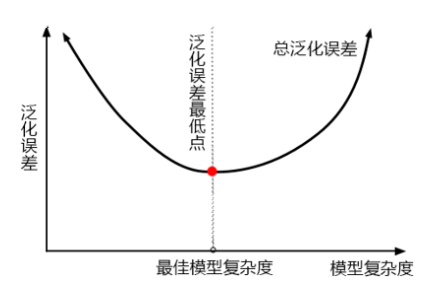
确认当n\_estimators=167时，达到该模型下最大准确度，停止对n\_estimators的调整

6.2调整随机森林中树的深度

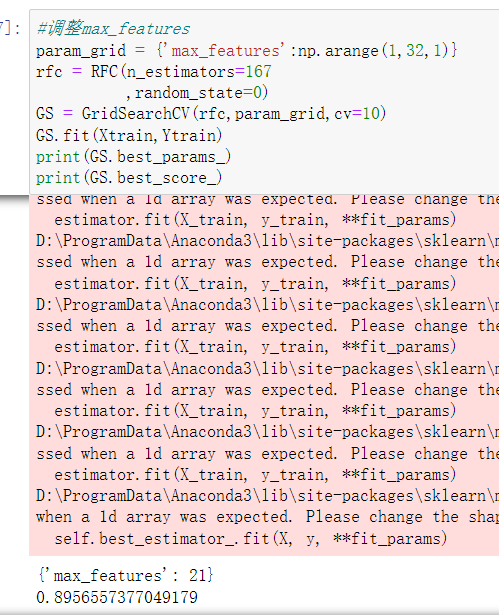
在n\_estimators=167的情况下利用网格搜索探寻树的最优深度



在这里，我们注意到，将max\_depth设置为有限之后，模型的准确率下降了。限制max\_depth，是让模型变得简单，把模型向左推，而模型整体的准确率下降了，即整体的泛化误差上升了，这说明模型现在位于图像左边，即泛 化误差最低点的左边（偏差为主导的一边）。通常来说，随机森林应该在泛化误差最低点的右边，树模型应该倾向 于过拟合，而不是拟合不足。这和数据集本身有关，但也有可能是我们调整的n\_estimators对于数据集来说太大， 因此将模型拉到泛化误差最低点去了。然而，既然我们追求最低泛化误差，那我们就保留这个n\_estimators，除非 有其他的因素，可以帮助我们达到更高的准确率。 当模型位于图像左边时，我们需要的是增加模型复杂度（增加方差，减少偏差）的选项，因此max\_depth应该尽量 大，min\_samples\_leaf和min\_samples\_split都应该尽量小。这几乎是在说明，除了max\_features，我们没有任何 参数可以调整了，因为max\_depth，min\_samples\_leaf和min\_samples\_split是剪枝参数，是减小复杂度的参数。 在这里，我们可以预言，我们已经非常接近模型的上限，模型很可能没有办法再进步了。 那我们这就来调整一下max\_features，看看模型如何变化。

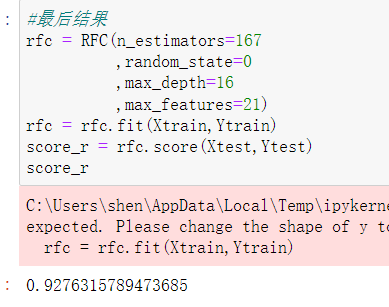


6.3调整随机森林中树的最佳分枝的特征个数

max\_features是唯一一个即能够将模型往左（低方差高偏差）推，也能够将模型往右（高方差低偏差）推的参数。我 们需要根据调参前，模型所在的位置（在泛化误差最低点的左边还是右边）来决定我们要将max\_features往哪边调。 现在模型位于图像左侧，我们需要的是更高的复杂度，因此我们应该把max\_features往更大的方向调整，可用的特征 越多，模型才会越复杂。

最终模型的准确度果然得到了提升，到了这里该模型已经达到了模型的预测上限，没有参数可以左右的部分了。剩下的那些误差,是噪声决定的，已经没有方差和偏差的舞台了。复杂度和泛化误差的关系已经告诉我们模型不能再进步 了。

模型最终的结果如下：



准确度为：0.9276315789473685