****

****

**QG中期考核详细报告书**

**题    目**

**学   院       计算机**

**专 业 计算机类**

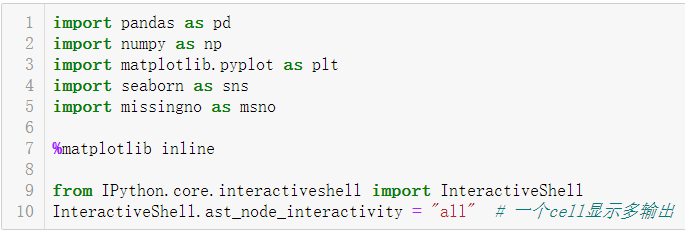
**年级班别 20级12班**

**学 号 3120005157**

**学生姓名 列校民**

**2021年 4月 16 日**

头文件引入



数据读入



首先使用pandas对数据进行读入的操作

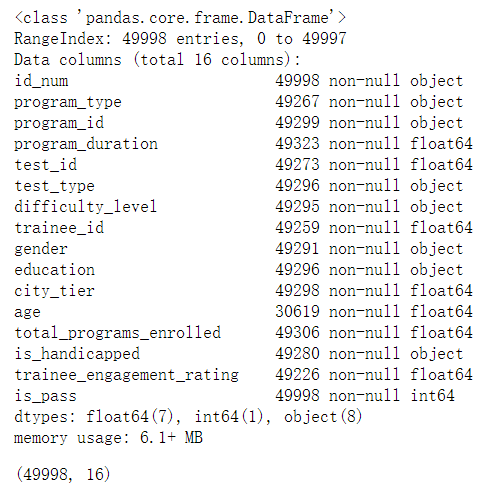
数据清洗

紧接着为了对需要处理的数据量有一个了解使用.shape的方法进行查看

再用info()方法与.shape得到的结果进行对比确定那些有缺失

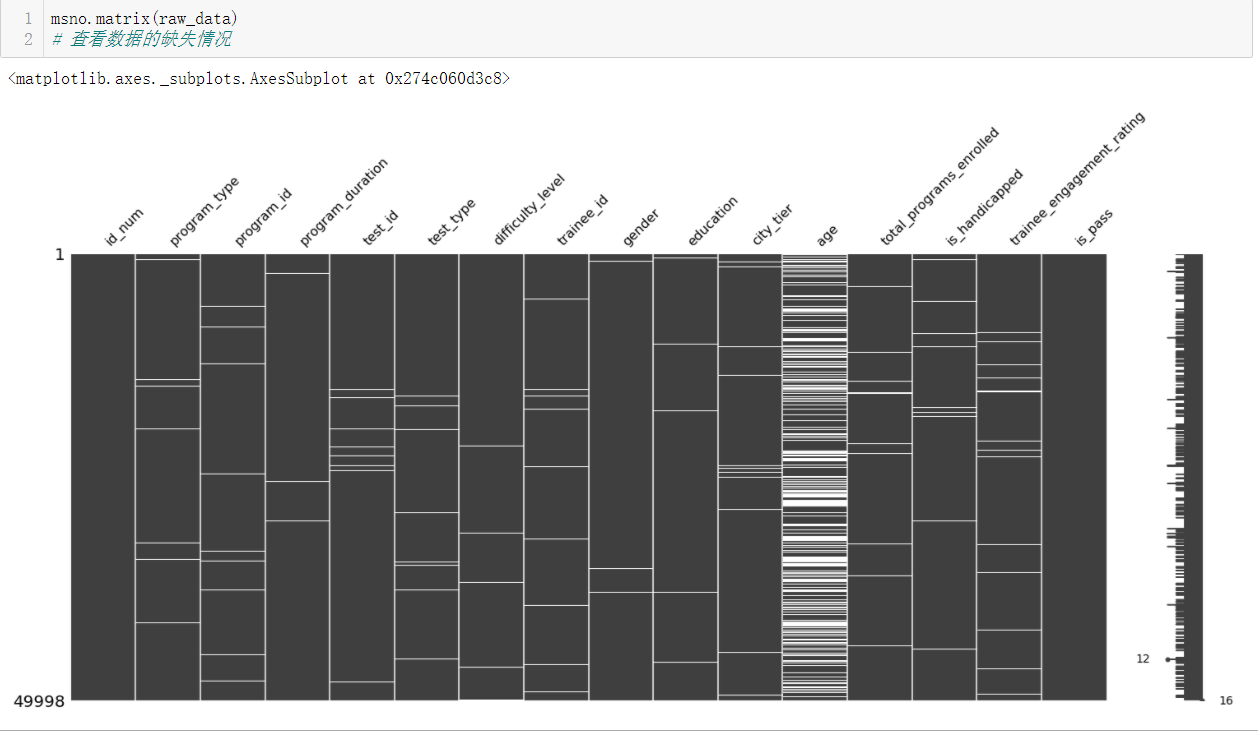
若某特征缺失的数据量占总的数据不大，则直接进行删除行数据的处理

若是某一列特征数据缺失明显数据量大，那么对其进行填充操作



但是发现这样对比非常缓慢而且不够直观

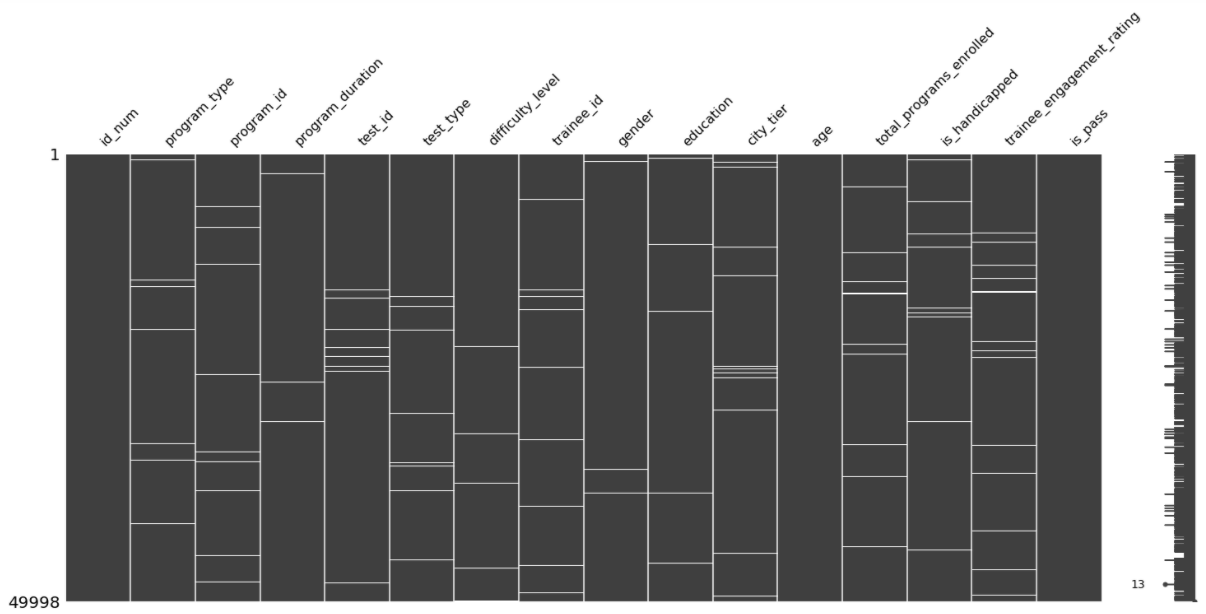
所以下面使用了图的形式对缺失的数据进行表示



观察得到age的数据缺失情况最为明显，所以我对age的数据进行填充平均值来剔除缺失值



再次构图查看填充是否完毕

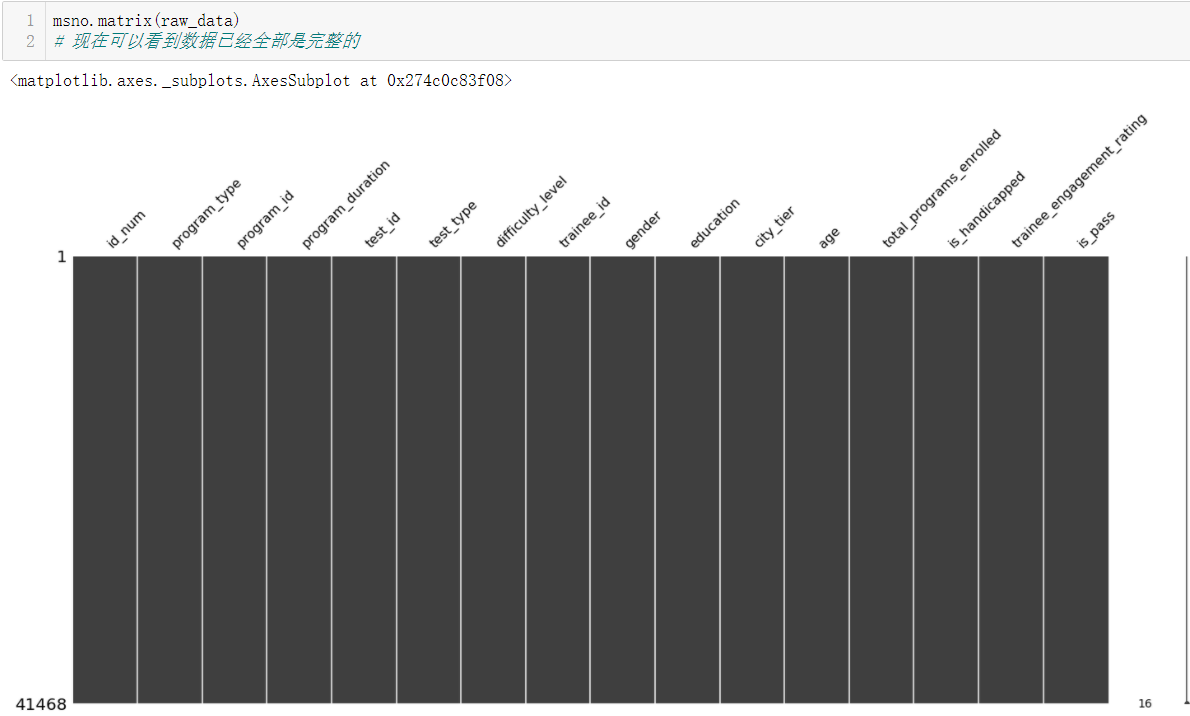


看到已经age填充完成，接下来继续对其他的数据进行操作

由于其他的数据缺失相对总数据较少，直接进行删除的操作，以免均值填充造成更大的误差

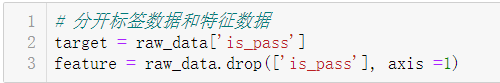
代码：





再进行可视化的查看，知道数据缺失的基本处理已经完成

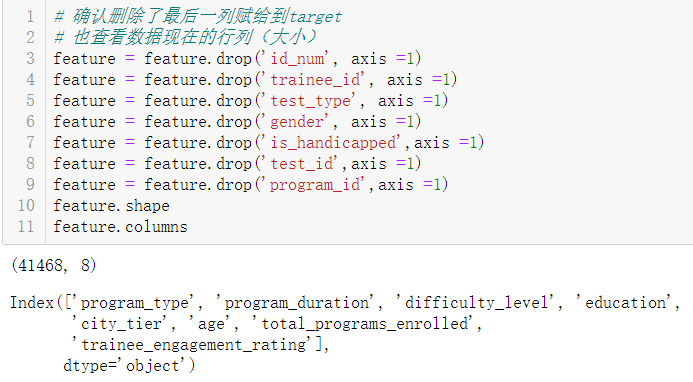
数据分集



删除明显无关的数据，对特征进行降维

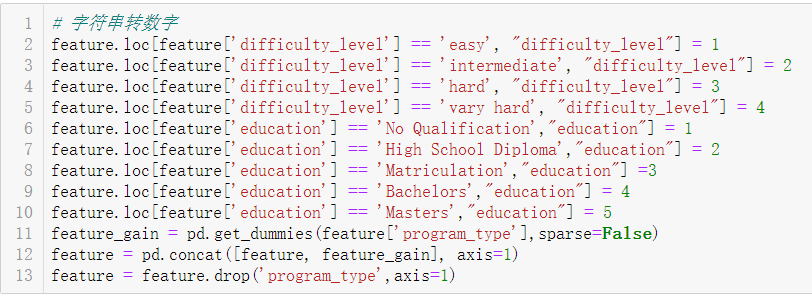
方差计算波动性





最后用feature.columns查看确认有真正清除不需要的数据

特征工程



将离散型数据转换

其中有高低等级之分的采用1、2、3、4的分级方法

而没有高低之分对的采用get\_dummies()函数进行one hot编码

之后再用concat进行数据的合并，最后删除掉已经被one hot编码的数据列

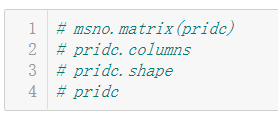
测试集的模型也同样操作

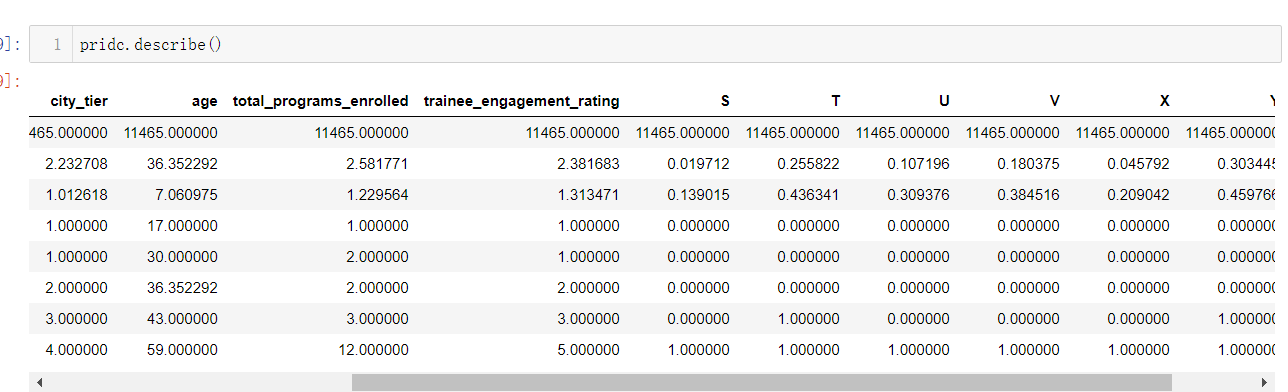




其实这里应该面向对象来实现会更好

一些对数据处理情况查看的反复调用cell





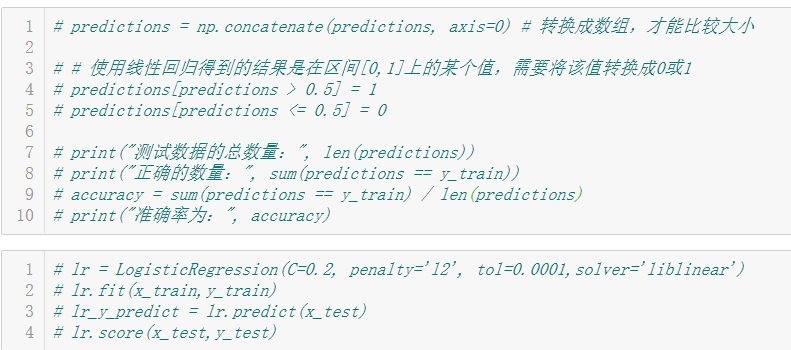
模型选择

调用三个模型，比较他们的准确率再确定模型





首先先进行对knn模型的训练得到较为合适的k值再与其他算法的模型进行比对



当然线性回归和逻辑回归也会调参，不会单独只进行对knn的调参

部分调参的说明

****c****：正则化系数λ的倒数，float类型，默认为1.0。必须是正浮点型数，越小的数值表示越强的正则化。

****liblinear****：使用了开源的liblinear库实现，内部使用了坐标轴下降法来迭代优化损失函数

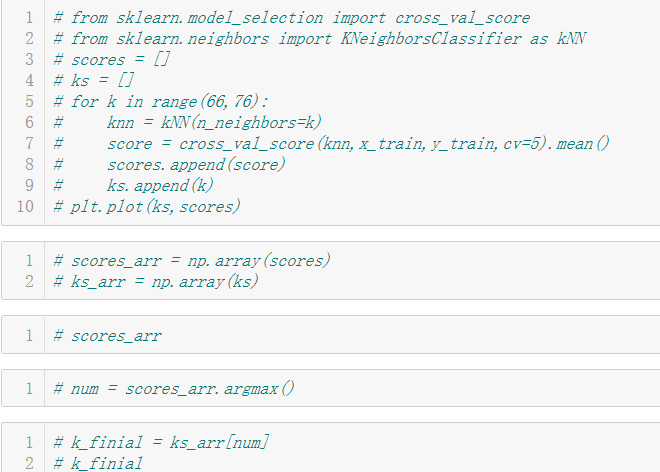
****solver****：优化算法选择参数，只有五个可选参数，即newton-cg,lbfgs,liblinear,sag,saga。默认为liblinear

****tol****：停止求解的标准，float类型，默认为1e-4。就是求解到多少的时候停止，认为已经求出最优解

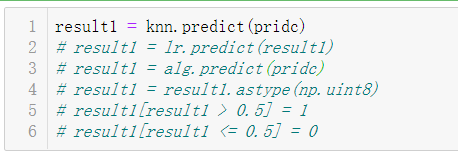
****penalty****：惩罚项，str类型，可选参数为l1和l2，默认为l2，这里的l1、l2为的其实是不让模型进行过拟合的正则操作

为什么用liblinear？liblinear适用于小数据集，而sag和saga适用于大数据集因为速度更快

Knn的模型训练



预测实现



写入文档

