****

****

**QG中期考核详细报告书**

**题    目     数据挖掘**

**学   院       计算机学院**

**专 业 计算机类**

**年级班别 20级计算机类10班**

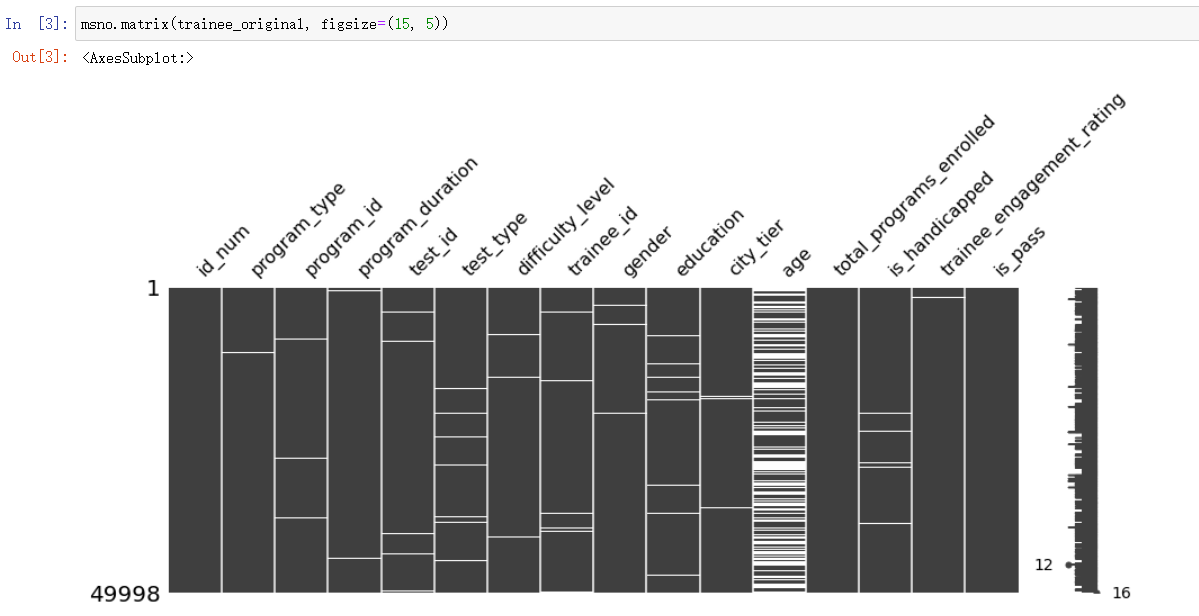
**学 号 3120005087**

**学生姓名 赵朗**

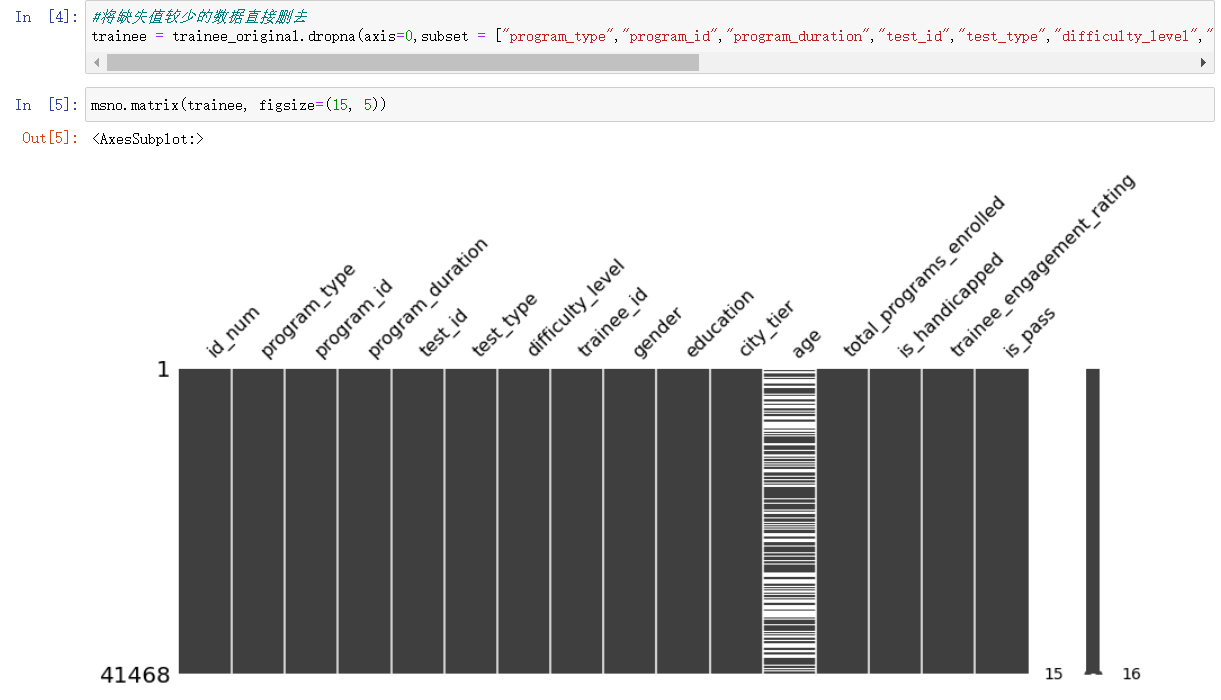
**2021年 4月 17日**

数据清洗：

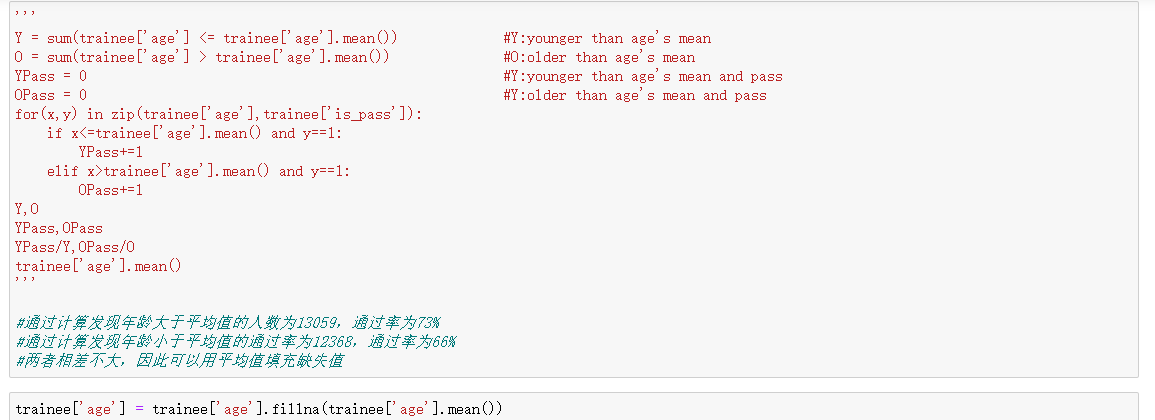
查看缺失值



除了age列，其余数据皆是零散的缺失，因此将有缺失值的数据直接删去



对含有缺失值较多的age列进行分析



因为年龄大于或小于平均值的人，他们的数量和通过率相差不大，因此用年龄的平均值来进行数据填充

数据分析：

通过观察发现program\_type, program\_id, program\_duration是互相对应的，为一类数据

因此根据不同program\_id的通过率对其进行赋值



计算test\_type特征不同特征值通过率并对其赋值



计算difficulty\_level特征不同特征值通过率并对其赋值



计算education特征不同特征值通过率并对其赋值



因为is\_handicapper中残疾的通过率为0，因此进行赋值

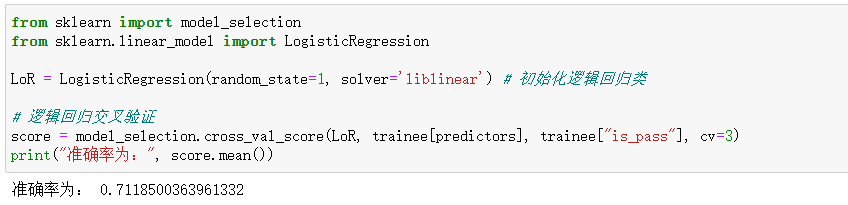


建立模型：

线性回归:调库并引用，建立线性回归模型，预测结果与实际结果比较获得准确率



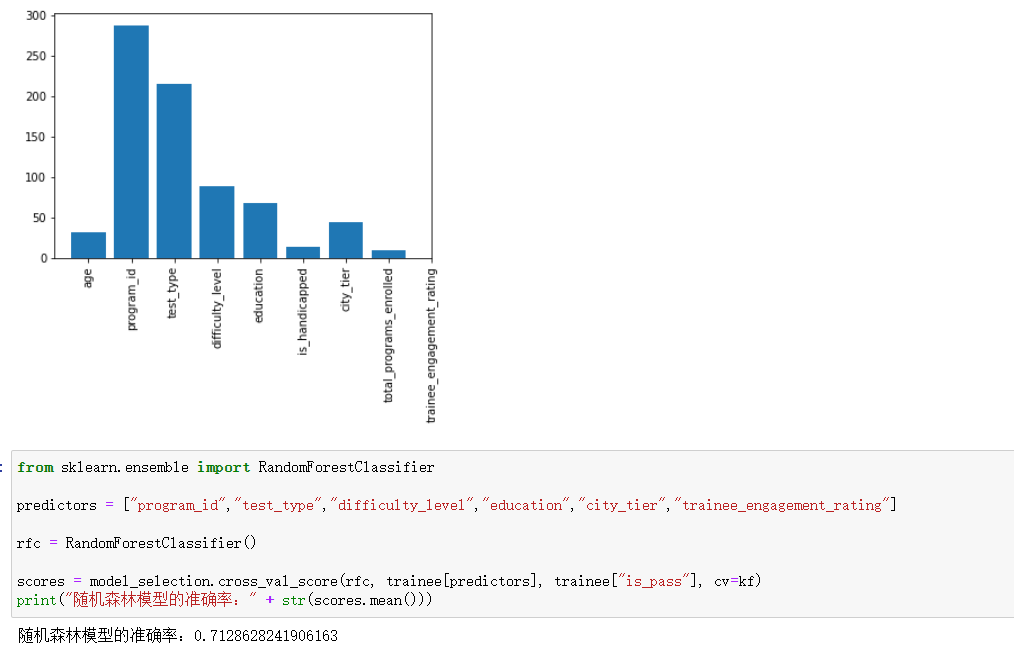
逻辑回归:调库并引用，建立逻辑回归模型，预测结果与实际结果比较获得准确率



梯度提升:通过调整max\_depth参数得到最大准确率的模型



特征选择：



模型选择:

根据不同模型准确率，来将准确率最高的模型应用在测试集中。