****

****

**QG中期考核详细报告书**

**题    目     中期考核数据挖掘过程**

**学   院       自动化学院**

**专 业 数据科学与大数据技术**

**年级班别 20级1班**

**学 号 3120001478**

**学生姓名 黄海涛**

**2021 年 4 月 16 日**

目录

[一、探明问题，理解数据 2](#_Toc69507669)

[（一）探明问题 2](#_Toc69507670)

[（二）理解数据 2](#_Toc69507671)

[二、数据预处理 3](#_Toc69507672)

[（一）数据清洗 3](#_Toc69507673)

[1、缺失值概览 3](#_Toc69507674)

[2、缺失值处理 5](#_Toc69507675)

[3、异常值检测 6](#_Toc69507676)

[三、特征工程 6](#_Toc69507677)

[（一）、特征删除 6](#_Toc69507678)

[（二）、特征处理 7](#_Toc69507679)

[1、归一化 7](#_Toc69507680)

[2、标签编码 7](#_Toc69507681)

[3、独热编码 7](#_Toc69507682)

[四、训练模型与模型优化 8](#_Toc69507683)

[（一）训练模型 8](#_Toc69507684)

[1、模型算法细节 8](#_Toc69507685)

[2、模型评估与选择 9](#_Toc69507686)

[（二）模型优化 11](#_Toc69507687)

[五、回顾与展望 11](#_Toc69507688)

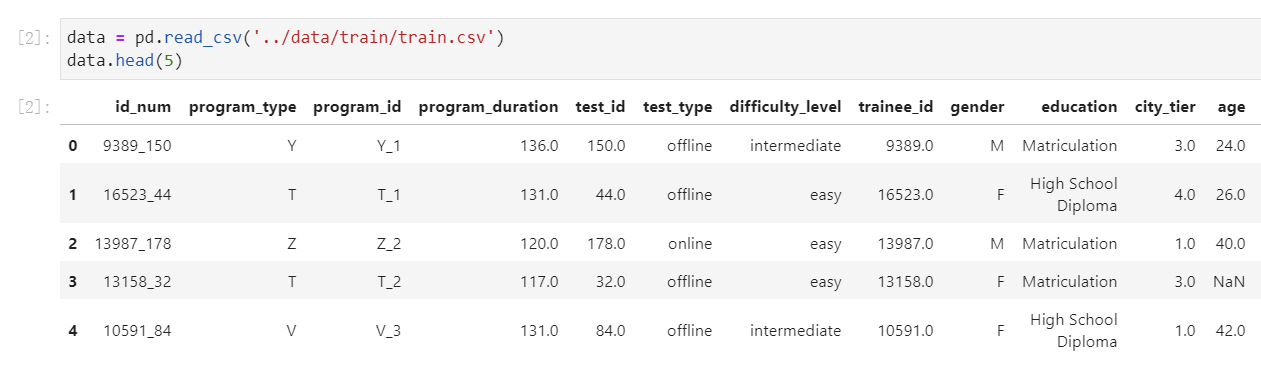
# 一、探明问题，理解数据

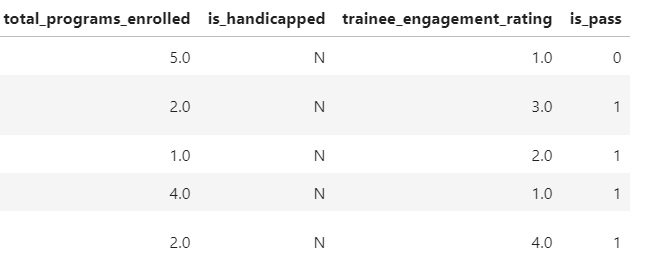
## （一）探明问题

由《中期考核数据集说明》可知，该份数据集为培训课程测试的**学员绩效**数据集。问题为通过找出最重要的因素来**提高受训者的参与度和表现**，这将使客户（培训公司）加强其培训问题。

## （二）理解数据

数据集共有三份，train为训练集，共有49998项学员数据。test1，test2为A榜和B榜对应的测试集，分别有11684项和11465项学员数据。通过pandas的read\_csv导入数据，将数据命名为data，调用dataframe的head方法，查看前5条数据。





根据Series的value\_counts可知，每个特征各有多少种值及统计情况。

现解释特征的含义

·id\_num，唯一ID，数据录入项凭证，是一个字符串型特征。

·program\_type，程序类型，是一个用于分类的字符串型特征。共有7种。

·program\_id，程序的ID，表明学员选择哪类程序下的哪个程序，是程序类型的进一步细分，是一个起到分类功能的字符串型特征。共有22项。

·program\_duration，程序计划持续时间，单位为天，是一个整型特征。

·test\_id，测试ID，学员选择的测试类型，一个用于分类的整型特征。共有188项。

·test\_type，测试类型（离线/在线），一个起分类功能的字符串型特征。共有2种值，为offline和online。offline为离线类型，online为在线类型。

·difficulty\_level，测试难度级别，分类型字符串特征。共有4种值，分别为easy，intermediate，hard，vary hard。

·trainee\_id，学员的ID，整形特征，共有16309项。

·gender，性别，分类型字符串特征，共有两种值，为M（male，男）或F（female，女）。

·education，学员的教育水平，分类型字符串特征。共有五种值，为No Qualification，High School Diploma, Matriculation，Bachelors, Masters。

·city\_tier，学员居住城市的等级，分类型整形特征。共有四种值，1，2，3，4。

·age，学员年龄，整形特征。

·total\_programs\_enrolled，学生注册的程序总数量，整形特征。

·is\_handicapped，受训者是否患有残疾，两种值，Y为患有残疾，N为不患有残疾。

·trainee\_engagement\_rating，讲师/教学助理为课程提供的学员参与度，序数型整形特征。

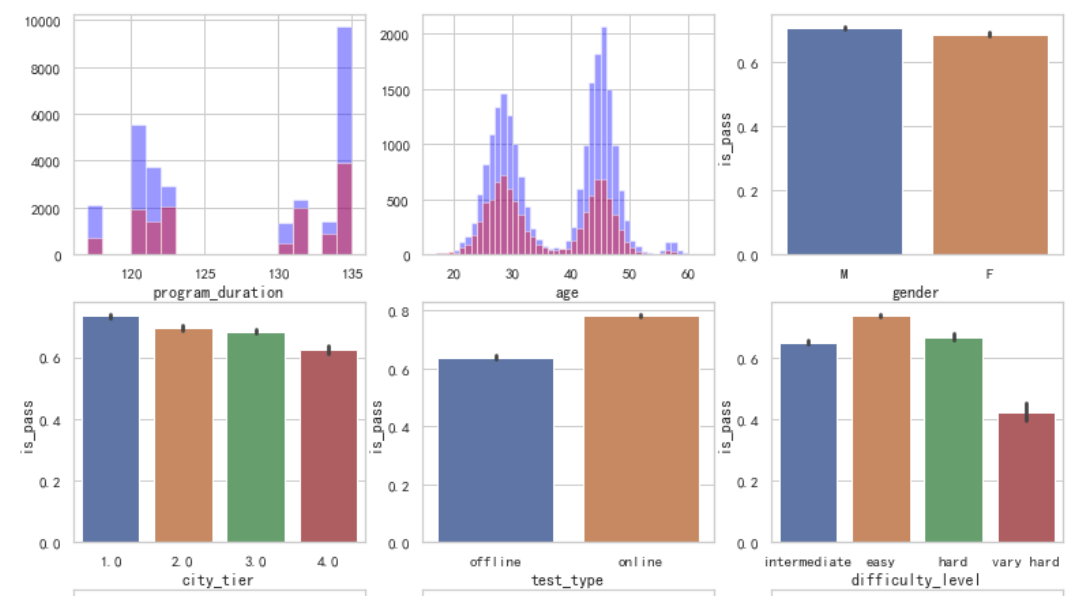
·is\_pass，0-测试失败，1-测试通过。

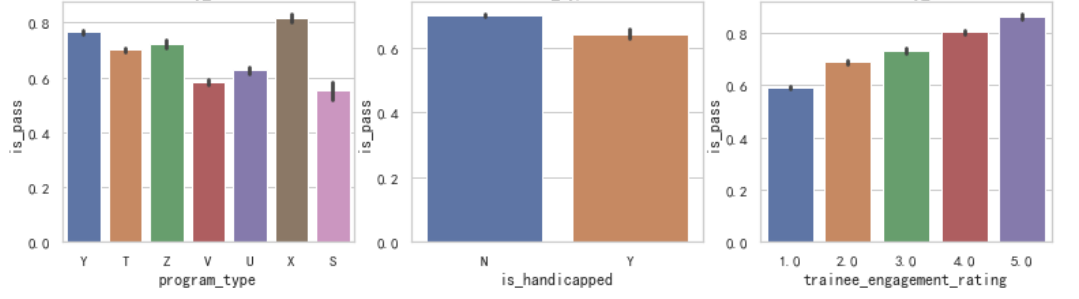
## （三）通过与未通过的分布比较

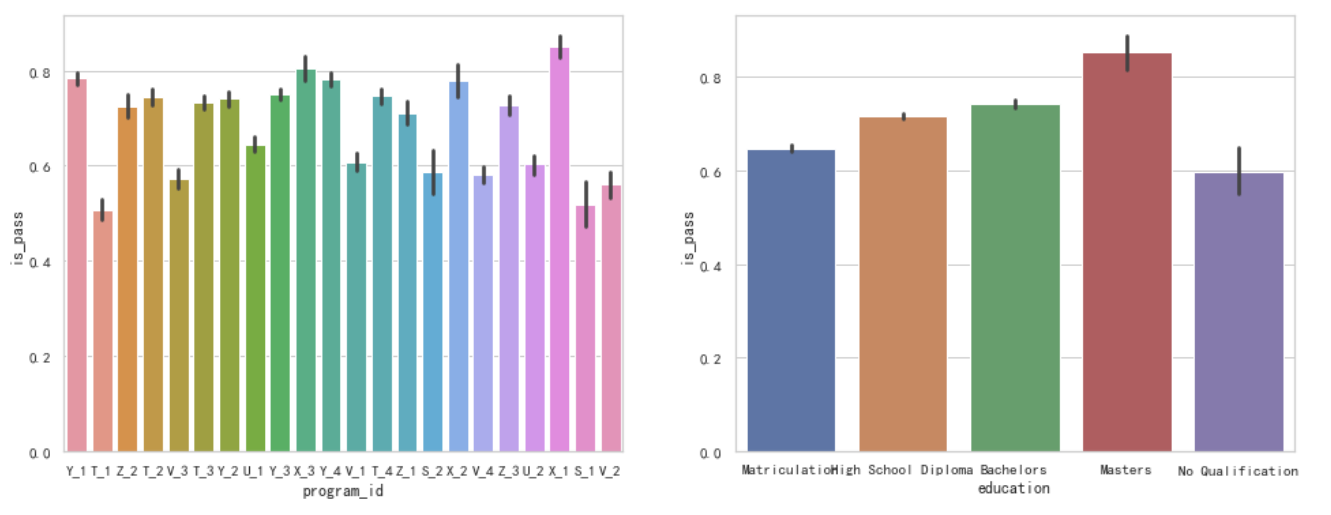
对于数值型特征，用蓝色标志通过，红色标志未通过，进行数量上的描绘

对于分类型特征，用通过率来描绘其分布。

以下图均采用seaborn第三方库绘制。







program\_duration135天的持续时间通过人数最多，但相应的未通过人数也是未通过中最多的。

age形成了两座山峰，年龄分布比较奇怪，可能是因为年龄特征在三十多岁左右有较大的缺失值，而我画的正好是去除过缺失值的，可能就正好去除了。

gender两座峰高相近，说明性别特征在is\_pass上没有区别。

city\_tier则体现了越不是一线城市，通过率越低，与我们的直观想法是吻合的。

test\_type，线上和线下在通过率上有着百分之十几的差距，相关性较强。

difficulty\_level，依照难度递增，通过率递减，符合直观想法。

program\_type和program\_id特征值间通过率上有着明显的差异。

is\_handicapped，两种特征值通过率相近。

trainee\_engagement\_rating，学员参与度越高，越容易通过，符合常理。

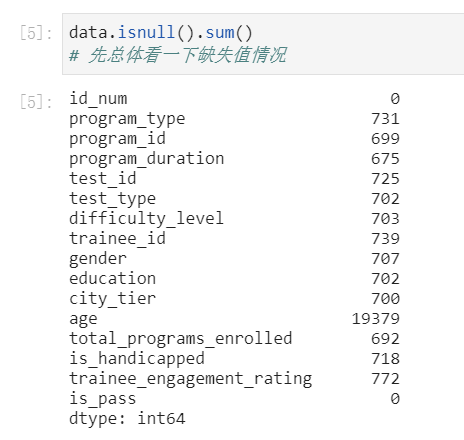
education，按照学历高低顺序，通过率依次递增，吻合平常的想法。

# 二、数据预处理

## （一）数据清洗

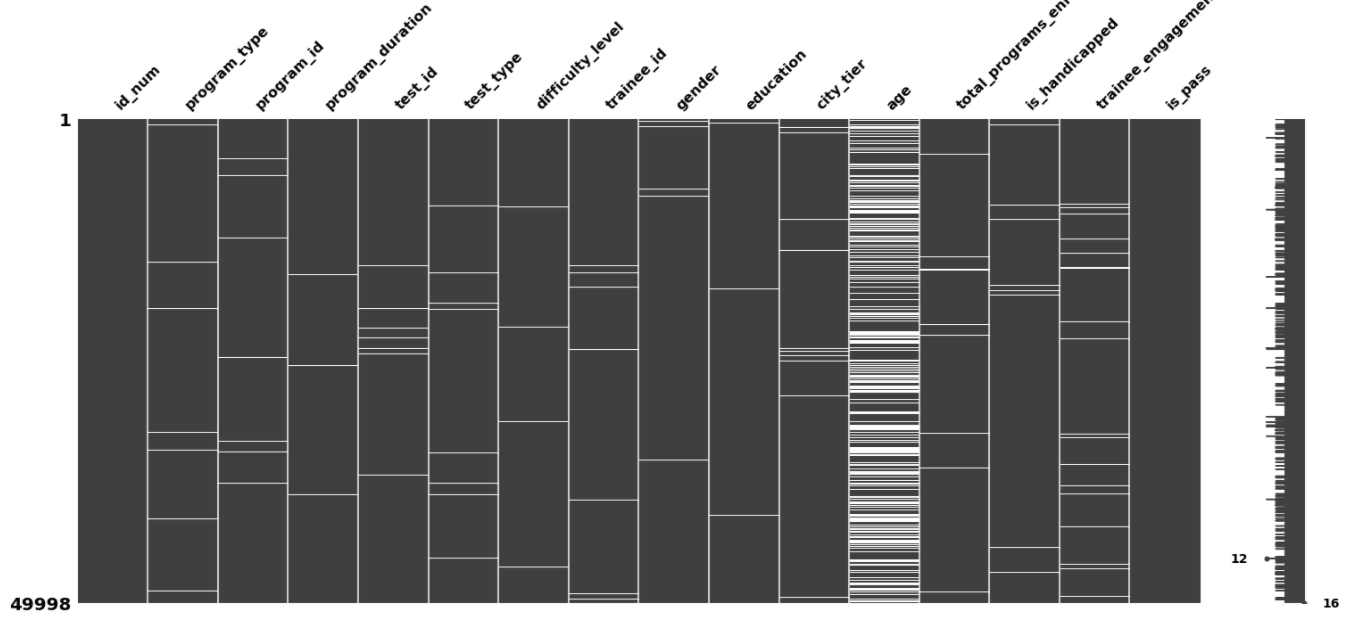
### 1、缺失值概览

data.isnull().sum()总体观察缺失值情况。



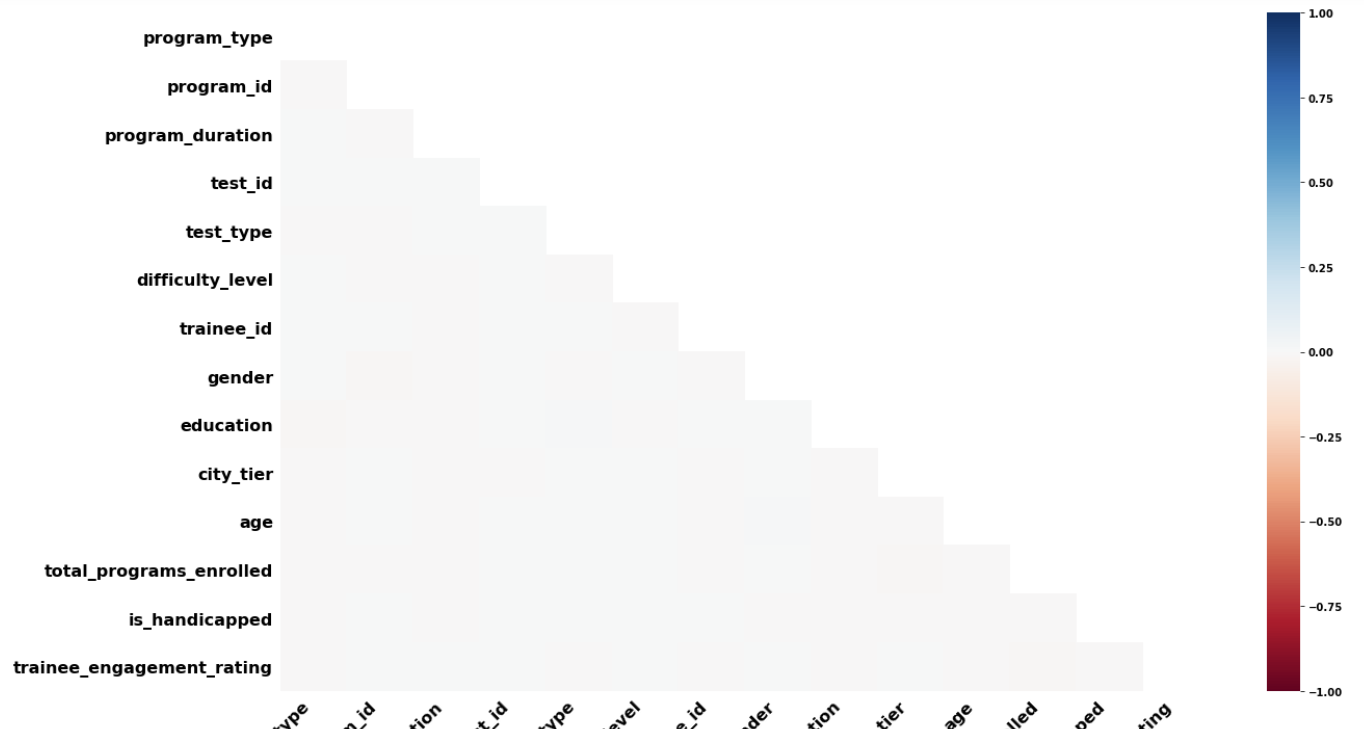
可见age特征近一半的数据缺失，其他特征缺失值比例小。

再利用missingno库的matrix方法可视化缺失值情况，得下图



**白线越多，缺失值越多**，再次说明了其余都是个别有缺失值，age则是大部分。

再利用missingno库的heatmap方法，查看缺失值出现情况是否有一定关联。



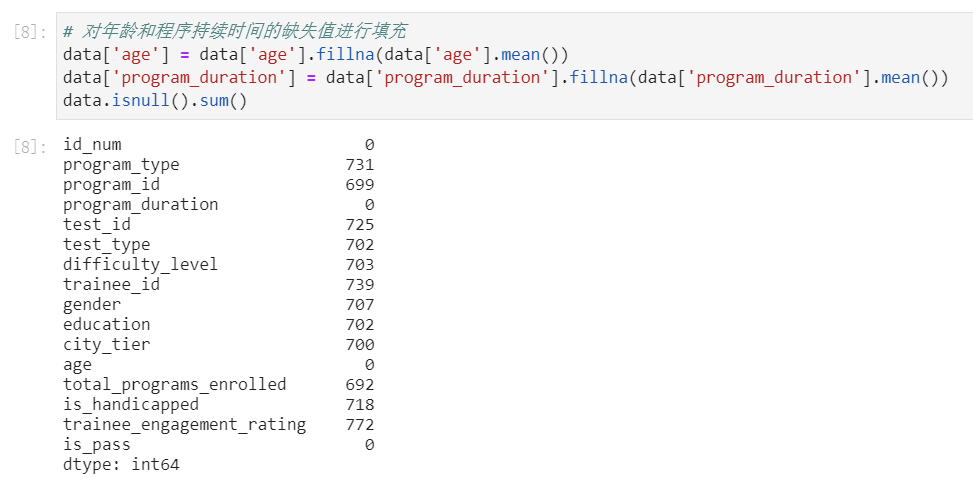
通过上图可知，缺失值间出现没有关联性。

### 2、缺失值处理

**缺失值类型如果为连续数值型特征，可以考虑用平均值或中位数或众数进行填充。如果是分类型特征，可以考虑直接删除。**

同时有另外一种思路，就是完全没有缺失值的数据来构建模型去预测缺失值进行填充，但做了一下发现并不可行。

连续数值型特征有age和program\_duration，用平均值进行填充。



分类型特征如果随意填充，模型效果会变差。

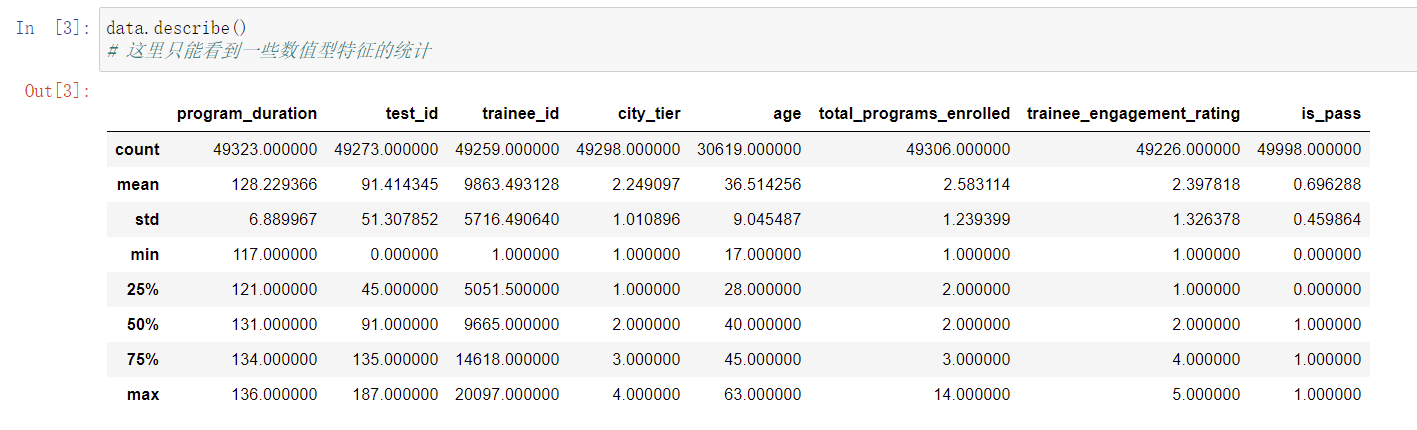
所以通过dropna()方法直接删除，共删去7954条数据，占总数据的六分之一。



### 3、异常值检测

通过dataframe的describe方法进行简单的描述性统计分析，可以先对变量做一个**描述性统计**，进而查看哪些数据是不合理的。最常用的统计量是最大值和最小值，用来判断这个变量的取值是否超出了合理的范围。

可见，这份数据集并没有异常值。



异常值检测的方法还有**3σ原则**，在正态分布中σ代表标准差,μ代表均值。x=μ即为图像的对称轴。数值分布在（μ-3σ,μ+3σ)中的概率为0.9974，因此可以认为分布在该范围外的数值为异常值。

但本人没有在本次分析中运用。

# 三、特征工程

## （一）、特征删除

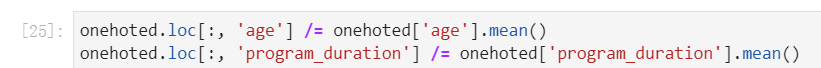
trainee\_id和test\_id为无序的离散值（虽说他们和is\_pass的相关性挺高的），id\_num为唯一ID，这三个特征都只是起到了身份凭证的作用，考虑将其直接删去。而programme\_type和programme\_id有一定的顺序规律，为分类型特征，但又因为programe\_id值也表明了programe\_type的身份（例Y\_1表明了他programe\_type为Y），所以可以删去programe\_type，留下programe\_id。

基于我们对数据的理解，我们对id\_num，program\_type， test\_id， trainee\_id进行删除。

## （二）、特征处理

### 1、归一化

将age和program\_duration两项特征进行归一化，这样可以**加快运行速度**。

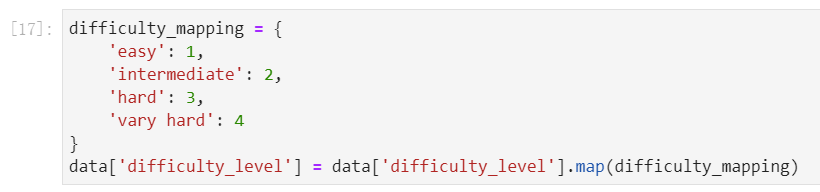


### 2、标签编码

**标签编码**是对特征进行转换，变成连续的数值型变量。即是对不连续的数字或者文本进行编号。

因为本次我要使用的是自己编写的逻辑回归算法，只接受数字运算，所以对于**二元分类**特征或**多元分类**下每种值间有**递进关系**的字符串型特征，要将其进行标签编码，转换成数字。对于test\_type，其有两种值offline和online，为二元分类特征，所以转换成0和1。对于difficulty\_level，其有四种值，easy，intermediate，hard，vary hard，值间有着明显的**递进关系，**所以对应转换成1，2，3，4。

而在转换过程中，我编写了**字典**描述对应关系，再将其作为series的map方法的参数对值进行标签编码。举一个例子：



这种字典的方式简化了进行标签编码的过程，且**直观优雅**。

对test\_type，gender，education，difficulty\_level，is\_handicapped进行了标签编码。

### 3、独热编码

特征删除时曾提到留下了program\_id，但在标签编码处没有对其进行处理，是因为它的值并没有像difficulty\_level那样有着递进关系，而是单纯的分类，如果用标签编码表示，如1，2，3，这样会隐含3比1重要的意思，但实际上并没有。所以，要用**独热编码**来表示这些没有递进关系的特征，可以理解为独热编码是根据特征数量创建相应维数，将每一类表示成一个单位向量，但每一个单位向量的方向不一样，从而就达到了同级不同类的效果，也避免了“3比1重要”这种对计算机的误导。

此处利用了pandas的get\_dummies方法将program\_id独热编码，然后用pandas的concat方法将独热编码后的特征矩阵与原特征矩阵拼接起来后再删去原program\_id列删去。

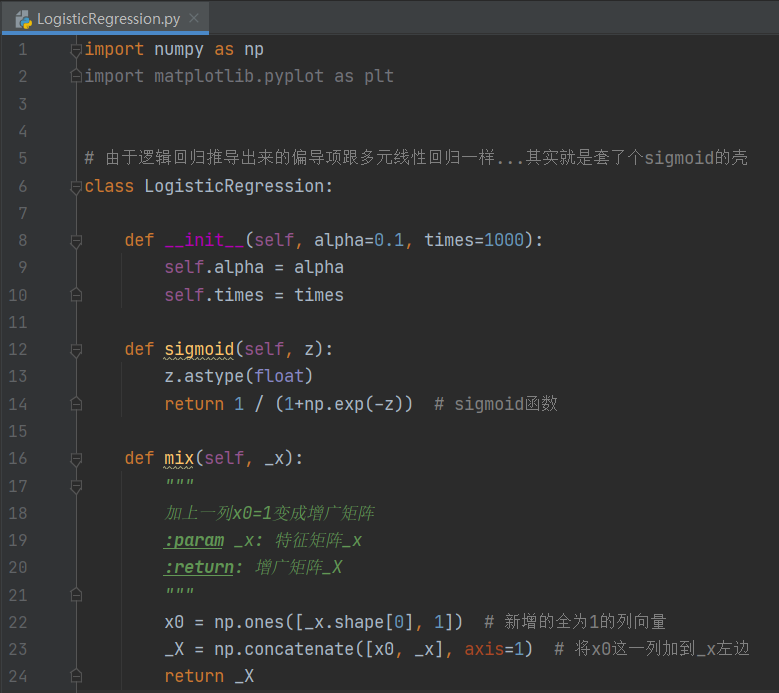


# 四、训练模型与模型优化

## （一）训练模型

做的是二元分类预测，所以使用了自己编写的LogisticRegression算法。在此展示其内部细节。

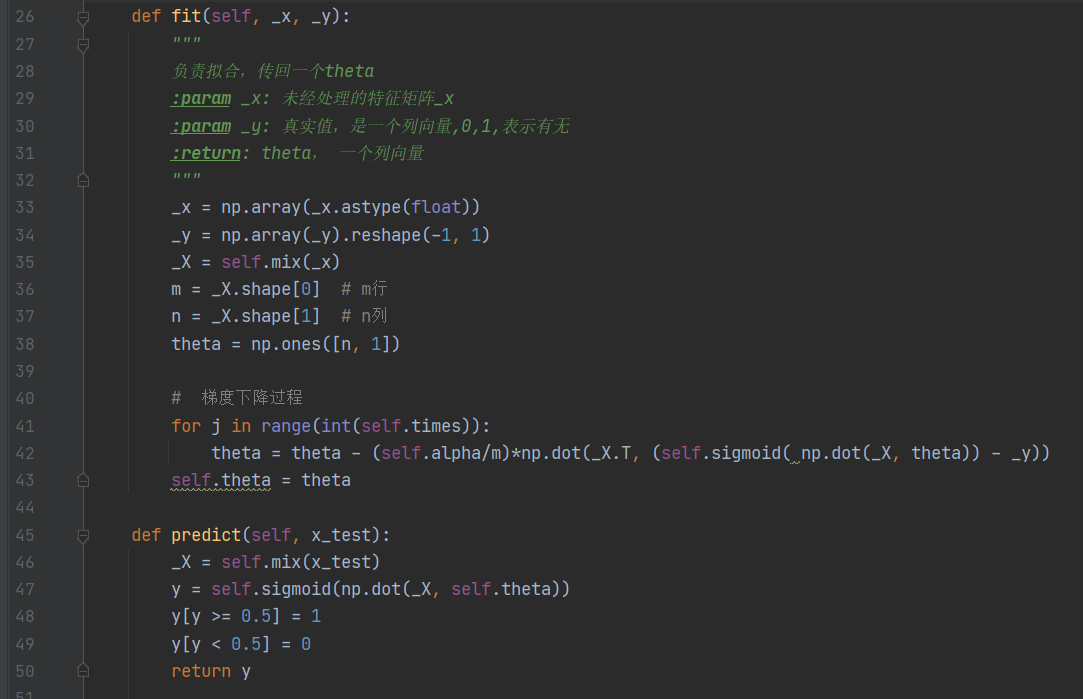
### 1、模型算法细节



\_\_init\_\_是类的初始化方法，设置一些参数。我这里默认学习率alpha=0.1，迭代次数times=1000。

sigmoid函数将输入的值映射到(0,1)内，变为一个概率值。

mix函数则是对传进来的矩阵增添一列x0，全部赋值为1，变为增广矩阵。目的是增加偏置项。



fit函数则是模型用于拟合的函数。这里的细节就不讲了。只讲两个地方。

1.为什么\_x要astype(float)？

因为pandas进行运算时很多函数都是基于数据类型为浮点型进行运算，int型会出一些奇奇怪怪的问题。

2.为什么要\_y要.reshape(-1,1)？

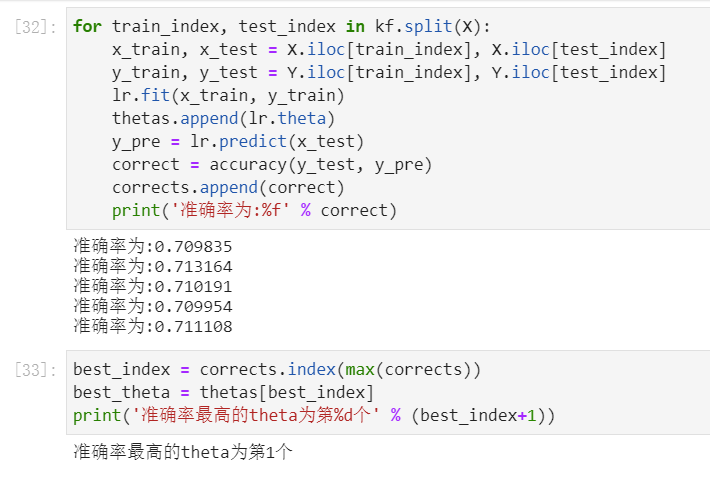
这是我在拟合乳腺癌数据集时发现的一个bug，当输入的y没有reshape时，（假设数据是m行数据n列），\_y的形状是(m,)，\_X与θ相乘得出预测值的形状是(m,1)，它们一相减就会**得出一个(m,m)的矩阵**。通过stackoverflow查找得知是numpy内对(m,)和(m,1)的处理是不一样的（虽说看起来一样），所以要对其指定形状，运用到reshape()，同时用到**-1**这个看起来摸不着头脑的参数，Numpy会根据剩下的维度算出另一个shape属性值。

### 2、模型评估与选择

导入sklearn的留一法和KFold类。用两者种分割数据集的方法进行测试，每次拟合把theta和准确率accuracy加入相应的列表，最后根据**准确率**来选择效果最好的theta。

代码如下：

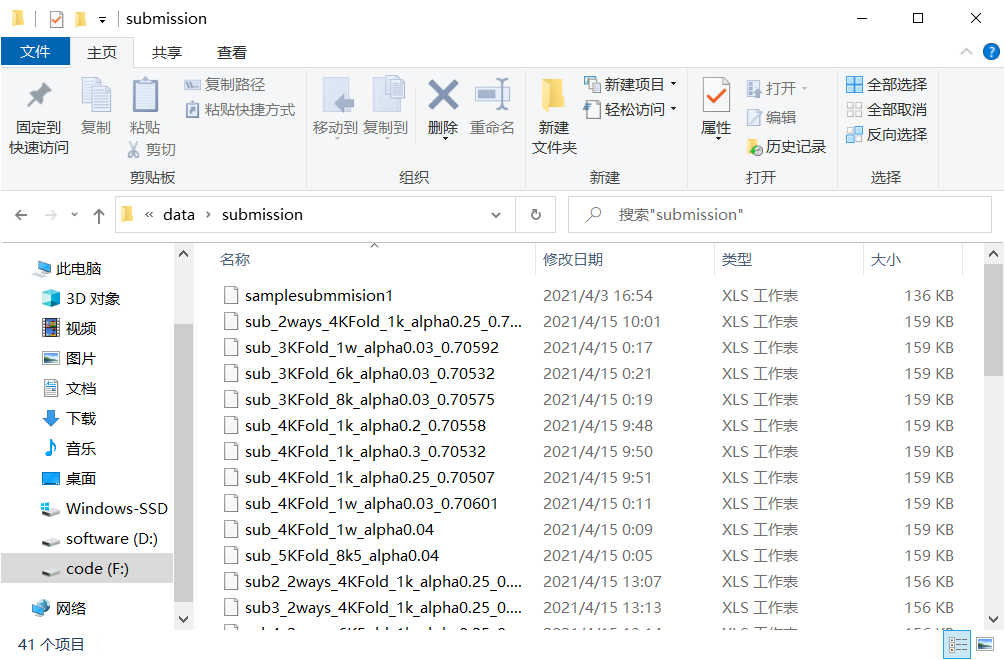




之后就将best\_theta赋给模型的theta属性，用这个模型去预测test数据（test的数据也要按拟合时那样进行相应的数据处理）。

## （二）模型优化

dataframe生成csv文件时，名字内有此次的参数，通过记录每一次提交的分数，对参数进行相应的修改，如留一法中的test\_size，KFold实例化时的n\_splits，LogisticRegression实例化时的学习率alpha，迭代次数times。



# 五、回顾与展望

通过此次中期考核，对pandas和numpy的函数熟悉了很多，也了解了很多数据处理，分析的方式并予以应用，也算是做了一次正式的数据挖掘，学到了很多。

不足：

1.特征上处理并不合理，虽说我通过相关系数矩阵和SelectKBest得知test\_id和trainee\_id与is\_pass有很强的相关性，但我还是删掉了它，毕竟我觉得id这类特征纯粹是巧合…

2.分析过程图不够，显得整个分析过程不够直观。如特征的密度图。

3.没有了解该图数据分布是否倾斜，如果倾斜可以采取log变换，更符合正态分布，并使得数据与分布的平均值无关。

4.算法知识的匮乏导致可选择模型少，准确率被限制在一定范围内。

自己的一点感想：

从无到有，需要勇气，这个走出舒适区的过程必定是痛苦的，但不迈出这一步就会错失提升自我的机会，也会一直恐惧变化，这并不利于个人的发展。所以，多尝试，多搜索，不会就问！干巴爹！冲冲冲！