

**QG中期考核详细报告书**

**题 目 数据挖掘报告**

**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2班**

**学 号 3220001560**

**学生姓名 周雍圆**

**2021 年 4 月 15 日**

目录

[1.题目分析 2](#_Toc31159)

[2.数据分析 3](#_Toc28027)

[3.数据清洗 6](#_Toc13306)

[3.1将非数字信息转化为数字信息 6](#_Toc12361)

[3.2缺失值处理 6](#_Toc27748)

[3.2.1关于缺失较少特征的处理 6](#_Toc939)

[3.2.2关于age的处理 6](#_Toc32510)

[4. 特征工程 7](#_Toc7614)

[4.1添加特征 7](#_Toc21184)

[4.2特征选择 7](#_Toc30119)

[5. 模型选择 9](#_Toc27737)

[5.1 KNN 9](#_Toc16573)

[5.2 线性回归 9](#_Toc22981)

[5.3 逻辑回归 9](#_Toc20982)

[5.4 随机森林 10](#_Toc30972)

[6.模型评估 10](#_Toc2253)

[7.学到了什么 10](#_Toc28573)

[8.个人感想 10](#_Toc18836)

**1.题目分析**

此处引用中期考核数据集说明。

“在多个程序中，您将获得针对培训课程测试明智的学员绩效数据集。您的任务是根据人口统计信息和培训计划/测试详细信息来**预测此类测试的性能**。通过**找出最重要的因素**来提高受训者的参与度和表现，这将使您的客户加强其培训问题。”

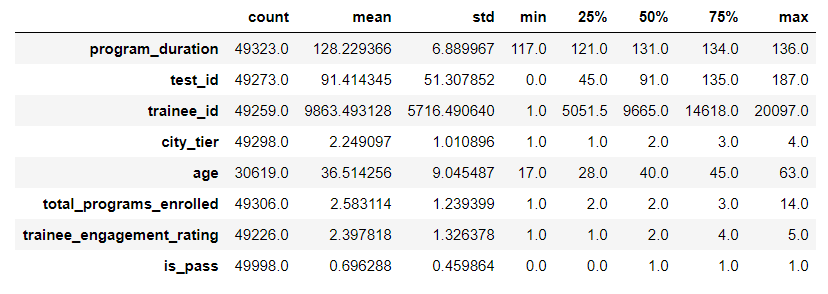
由此可见，该次考核有两个任务：

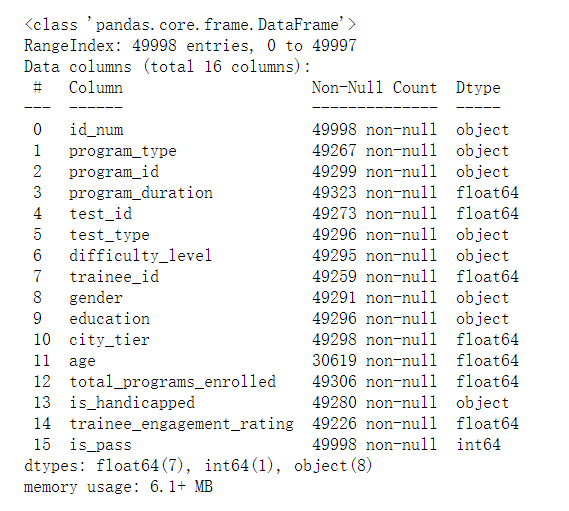
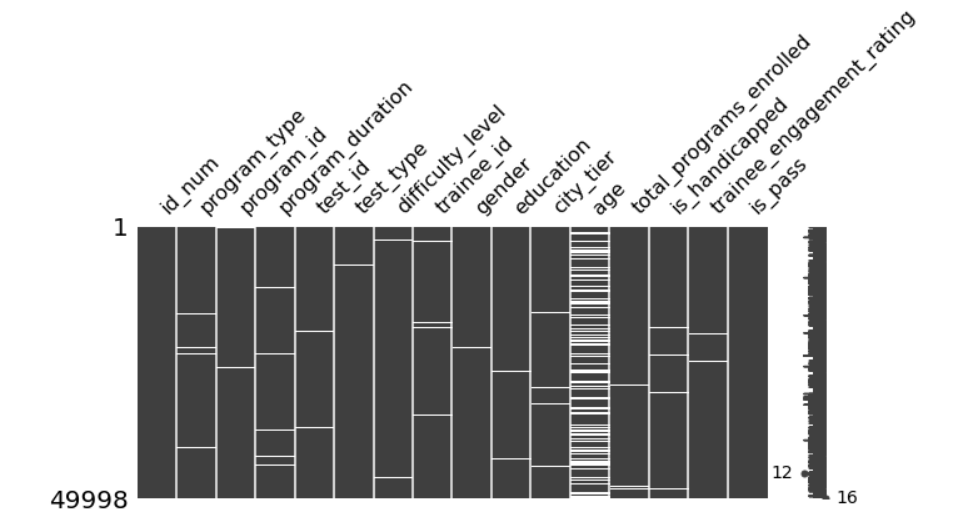
1. 找出提高受训者参与度和表现得最重要因素
2. 预测受训者的表现

任务一将在特征工程里的特征选择里进行回答。

**2.数据分析**

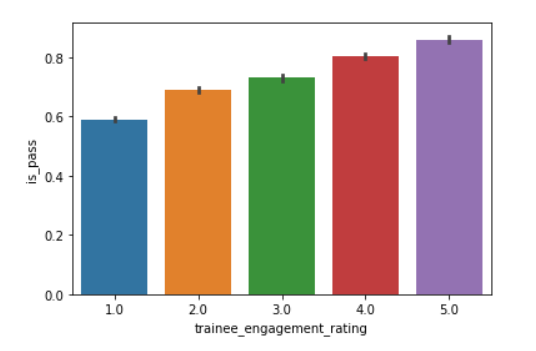
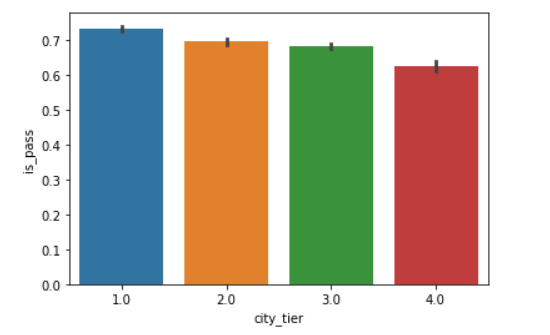
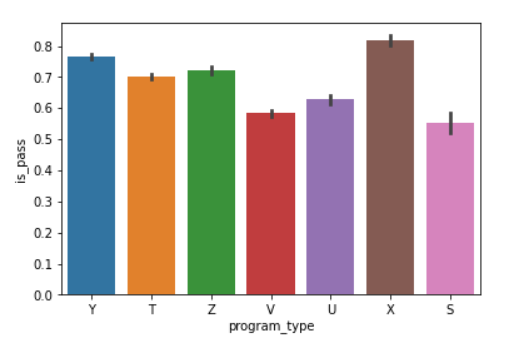
该训练集给出了16个特征，分别是id\_num,program\_id,program\_type, program\_duration,test\_id,test\_type,difficulty\_level,trainee\_id,gender,education, city\_tier,age,total\_programs\_enrolled, is\_handicapped, trainee\_engagement\_rating, is\_pass.

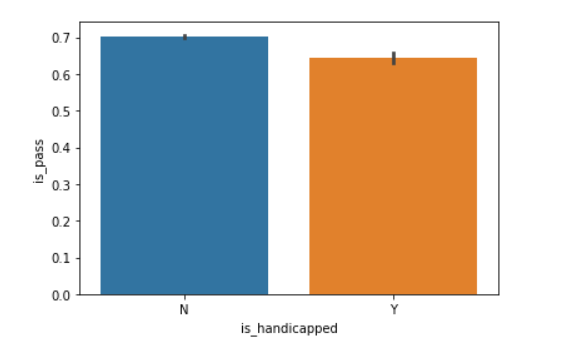
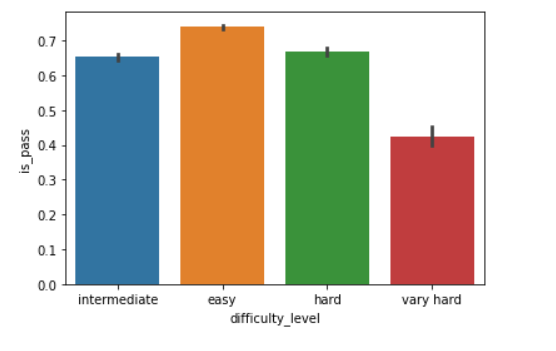


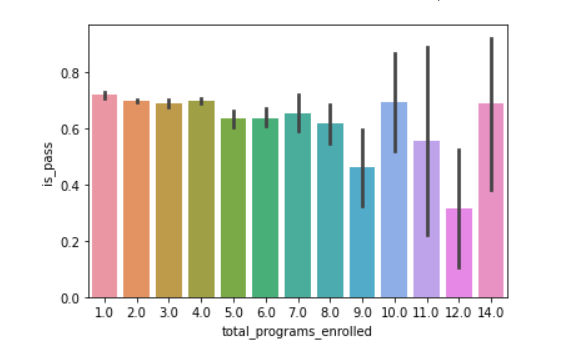


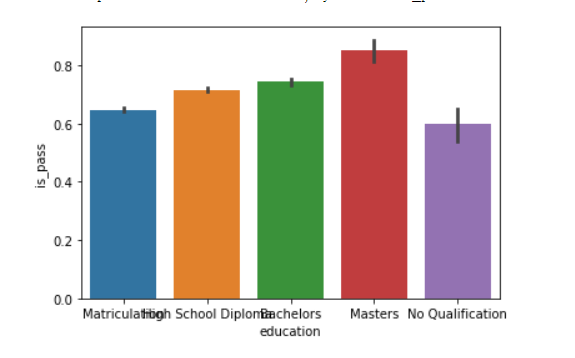
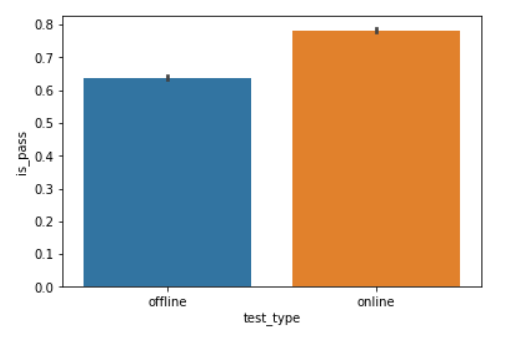
用pandas自带的describe()和info()查看基本信息，通过观察，发现数据有缺失值，暂时没有发现异常值。用missingno将数据可视化，可清洗看见哪个特征有缺失值。

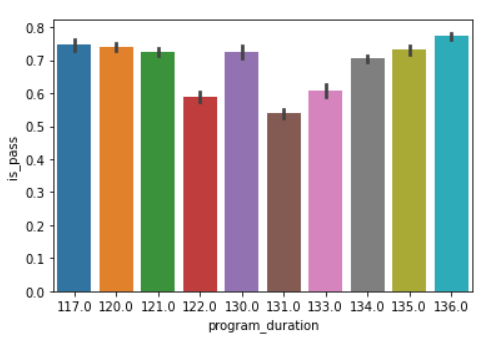
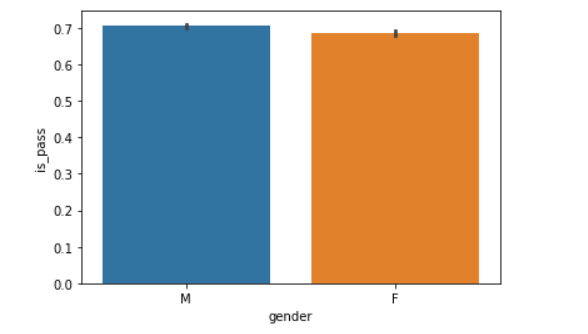
用seaborn画出特征与’is\_pass’的条状图。











这里没有选择id\_num, program\_id, test\_id, trainee\_id这四个特征。由图可以大致看出各个特征对应的通过率，其中比较明显的是，通过率随教育程度的增加而上升，随学员参与度的增加而上升。

除了id外的特征我全部都选择了，即program\_type, program\_duration,test\_id,test\_type,difficulty\_level,gender,education,city\_tier,age,total\_programs\_enrolled, is\_handicapped, trainee\_engagement\_rating。

**3.数据清洗**

**3.1将非数字信息转化为数字信息**

这里用的是独热编码（one-hot encording）。

一开始的时候我没有用连续编码，经继元师兄提醒后，将难度、教育程度都用连续编码了。

**3.2****缺失值处理**

以下是我曾经做过的操作。

**3.2.1关于缺失较少特征的处理**

1)使用众数填充

2)将其删除

由测试知将其删除准确率较高，最终选择将其删除。

**3.2.2关于age的处理**

1)使用平均值填充

2)用线性回归模型拟合已知年龄，进行对未知年龄的预测

3)用随机森林模型拟合已知年龄，进行对未知年龄的预测

由测试知随机森林模型拟合准确率较高，最终选择使用随机森林模型。

最开始的时候，我选择除id，age外的所有特征来预测年龄，利用了QG考核网站的检测准确率，每次删除一个模型，由此测试出用于预测年龄的最佳特征。这种方法不靠谱，个人认为随机程度较大。

最终选择了以下特征用于预测年龄：

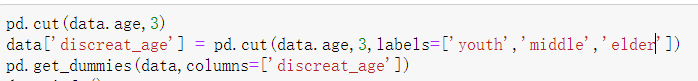
"program\_type","test\_type","gender","city\_tier","trainee\_engagement\_rating","program\_duration"

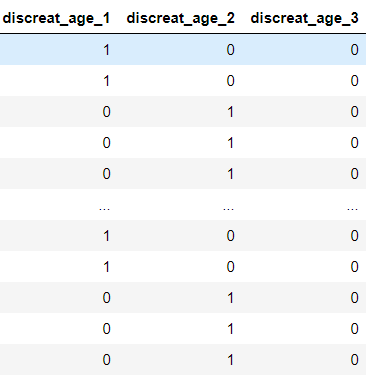
1. **特征工程**

**4.1添加特征**

想法：将年龄离散化，比如分出几个年龄段。

分出了三个年龄段，然后对此进行独热编码，瞎搞了一下。





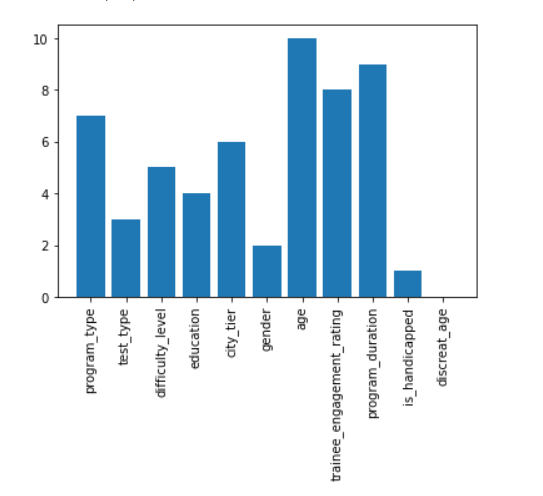
测试了一下，准确率略有微小提升，但可能是偶然。然后发现单独可以用，整体用会出BUG，故把它引了。

**4.2特征选择**

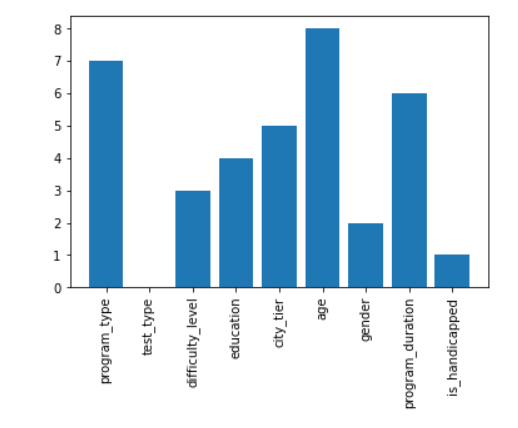
目前有想法但未实施：

用信息增益比筛选出对预测结果贡献最大的特征。

# 下图是使用sklearn中的RFE(Recursive feature elimination),即特征递归消除法，以选用特征为x\_train,‘is\_pass'为y\_train画图。纵坐标是排名，特征越优秀排名数值越大。下图我们可以看到影响通过率的比较重要的因素有age, program\_duration,trainee\_engagement\_rating.由于年龄这个特征很大部分是使用填充的，因为在此不选它作为最重要的因素。

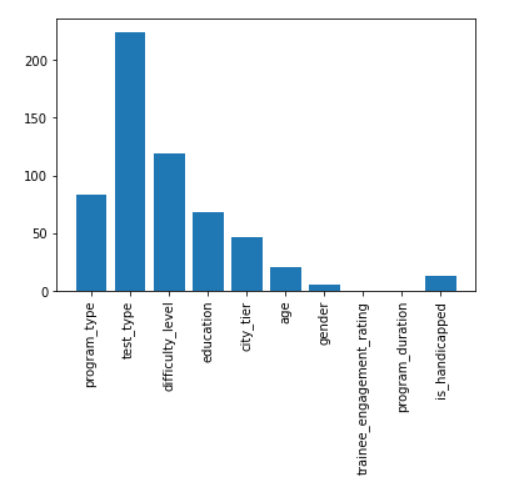


因为任务2)是找出提高受训者参与度和表现的最重要因素，因此将除trainee\_engagement\_rating外的选用特征作为x\_train，trainee\_engagement\_rating作为y\_train再画了一张图。同样，年龄特征因为大部分为填充所以不作选择。结合两图最终选择program\_duration作为提高受训者参与度和表现的最重要因素。



以上分析过程可能多少有点问题，做得还是相对粗糙。因为这个是根据随机森林模型选出最贴合’is\_pass’的特征。

CtrlC+V过继元师兄给的示范代码，画出来的图是这样的。



因为对模型原理不太熟悉，而却这个图看上去怪怪的，所以我没有采用这个方法。为什么说它怪怪的呢，因为在前面用seaborn画的图中trainee\_engagement\_rating明显是个重要的影响因素，然而在这张图中它的得分很低。也可能是我不清楚原理。

1. **模型选择**

以下是我曾经尝试过的模型。

* 1. **KNN**

原理：如果一个样本在特征空间中的K个最相似（即特征空间中最邻近）的样本中的大多数属于某一个类别，则该样本也属于这个类别。

使用sklearn中的KNeighborsClassifier，输入训练集，然后进行预测。

该模型为准确率最低的一个，而且耗时较长。

* 1. **线性回归**

用sklearn中的LinearRegression，用K折交叉验证调参。将预测值大于等于0.5的归为1，即为通过，小于0.5的归为0，即不通过。

* 1. **逻辑回归**

使用逻辑回归之前做了一下标准化，要不然会提示无法收敛。用sklearn中的LogisticRegression，输入训练集，然后进行预测。用交叉验证调参。

* 1. **随机森林**

使用随机森林的契机的看到了继元师兄的CSDN的例子，用的就是随机森林。没有太深究就import了，然后正确率就莫名提高了，害怕得我连夜回去学决策树。

原理：多个决策树组成随机森林，通过随机有放回的抽样，做到数据随机，特征随机。然后民主投票选出最好的选择。

用sklearn中的RandomForestClassifier，输入训练集，然后进行预测，进行交叉验证。

**6.模型评估**

1）准确率

显然所有的模型都用了准确率进行评估，即分类正确的样本占总样个数的比例。此处由QG测试网站提供的服务。随机森林准确率最高，逻辑回归准确率第二，线性回归准确率第三，KNN准确率最低。

**7.学到了什么**

1)numpy的基本功能，只有numpy部分看了书《数据科学手册》

2)决策树原理（只看了ID3和C4.5，CART还没看）

3)大致了解数据挖掘的流程

4)一点点，细微的统计知识。大致知道了几个名词，例如方差分析、F检测、卡方检测、T检测。即使简单看了下b站介绍的视频，但我还是没懂。

5)缺失值的处理方法（中位数填充、随机填充、向前/后填充、模型填充、删除）

6）时间管理

**8.个人感想**

作为一个零基础萌新，面对这种短期地爆发任务，压力还是很大的。首先是知识储备不足，三件套才飞速看了一点，机器学习方法也只学了3种，然后上路了，在实践中学习。我看了挺多CSDN，做了很多尝试，参考着别人的代码，照葫芦画瓢，跌跌撞撞地走到了这里。

我认为，我不是一个学得快的人。我学习比较需要时间，压缩时间飞快地看一遍的话，估计过段时间什么都记不住。这段考核时间段里，我对自己还是不太满意的。

首先是心态。排行榜这个东西确实有种无形的压力，看到自己的排名刷刷往下掉多少有点心态失衡，结果就是功利地套了随机森林的模型，准确率确实是提升了，但是自己没有收获什么，有段时间觉得自己是在为了冲榜而冲榜，不是在考核中学习，很麻木地在改代码。

然后就是学习扎实度。迫于时间压力，我确实有点急于求成，很多原理理解得根本不够透彻，就开始运用。可能就是看几篇CSDN，看个短视频，大致了解了是用来干嘛的就开始用了，更深层的原理、公式基本一概不知。

所以在昨天晚上，我停下来了，没有再去冲榜。我认真地看了一下《统计学习方法》里面决策树的部分，看了一些统计知识（虽然没有看懂），但是感觉好多了。把基础学扎实了确实比较重要。

最后，在做一样东西的时候，最开始不要总是想着要做到完美，主要是先做出来，然后再一个想法一个想法地试，再优化。接触新的东西确实很痛苦，同时也挺有趣，看着自己地准确率一次次的提升，偶尔也是会有成就感的。