****

****

**QG中期考核详细报告书**

**题    目     通过测试吧！!**

**学   院      计算机学院**

**专 业 计算机类**

**年级班别 20级4班**

**学 号 3120004781**

**学生姓名 陈劲铧**

**2021 年 4 月13 日**

目录

1. **理解题目·················3**
2. **分析数据·················4**

**三、对数据进行预处理，特征工程········5**

**四、尝试算法，模型优化，调整参数·······8**

**五、导出成果·················12**

**六、不足之处·················12**

**一、理解题目**

## 数据描述  
您的客户是一家跨国金融公司，向消费者提供多种产品。尽管主要的贡献来自线下分销渠道，但是有多个渠道可以为消费者提供这些产品。离线渠道通过其代理商网络向消费者出售金融产品，并且根据政府规定，这些代理商必须获得销售金融产品的认证。针对不同类别的金融产品，有多种认证计划。  
  
由于这个离线渠道对公司的总销售额贡献很大，因此公司着重于招聘并认证他们以建立大型代理商网络。在这里，主要的挑战是**培训他们获得销售各种类型产品的认证。**  
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。**  
## 内容  
- 变量-说明  
- id\_num-唯一ID  
- program\_id-程序的ID  
- program\_type-程序类型  
- program\_duration-计划持续时间（天）  
- test\_id-测试ID  
- test\_type-测试类型（离线/在线）  
- 难度级别-测试难度级别  
- trainee\_id-学员的ID  
- 性别-受训者性别  
- 教育-学员的教育水平  
- city\_tier-实习生居住城市的等级  
- 年龄-受训者年龄  
- 总计划招收-总课程的学生通过实习  
- is\_handicapped-受训者是否患有残疾？  
- 学员参与度-讲师/教学助理为课程提供学员参与度  
- is\_pass 0-测试失败，1-测试通过

从上文可以提出一下关键点

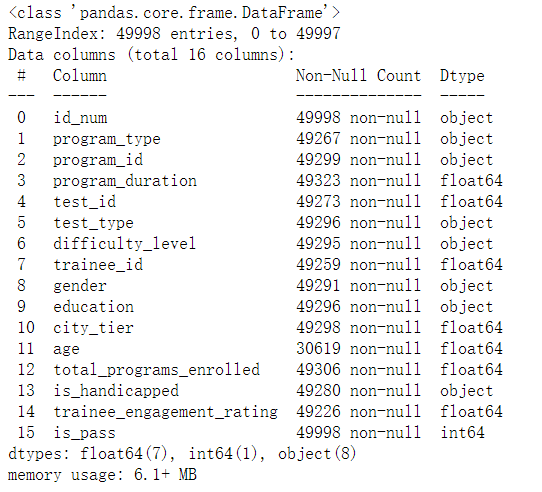
1. 在训练集中
2. 通过找出最重要的因素
3. 提高受训者的参与度和表现，进而使您的客户加强其培训问题

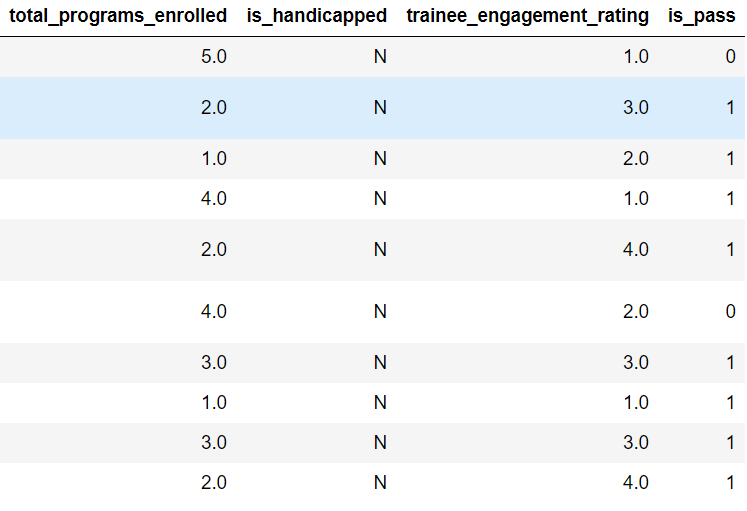
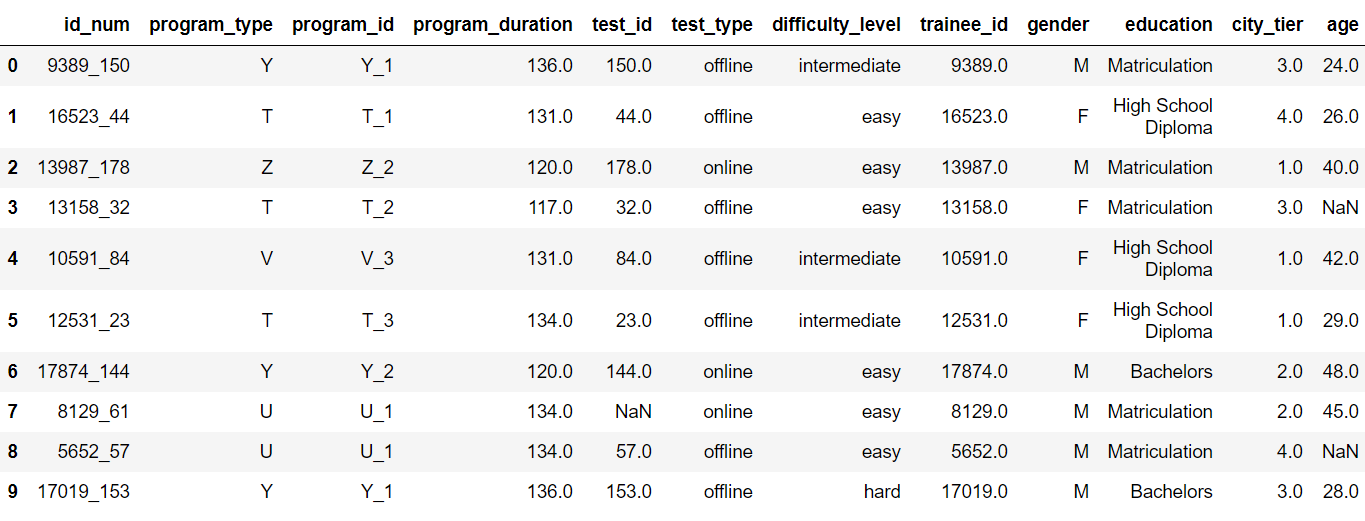
而我们所要做的：找出**最重要的因素**！！

而我在想，如何表示重要性，想到了**权重。**

**然后发现自己并不知道如何表示权重，想说待会再看看吧，先跑跑模型，估摸一下数值。**

**二、分析数据**



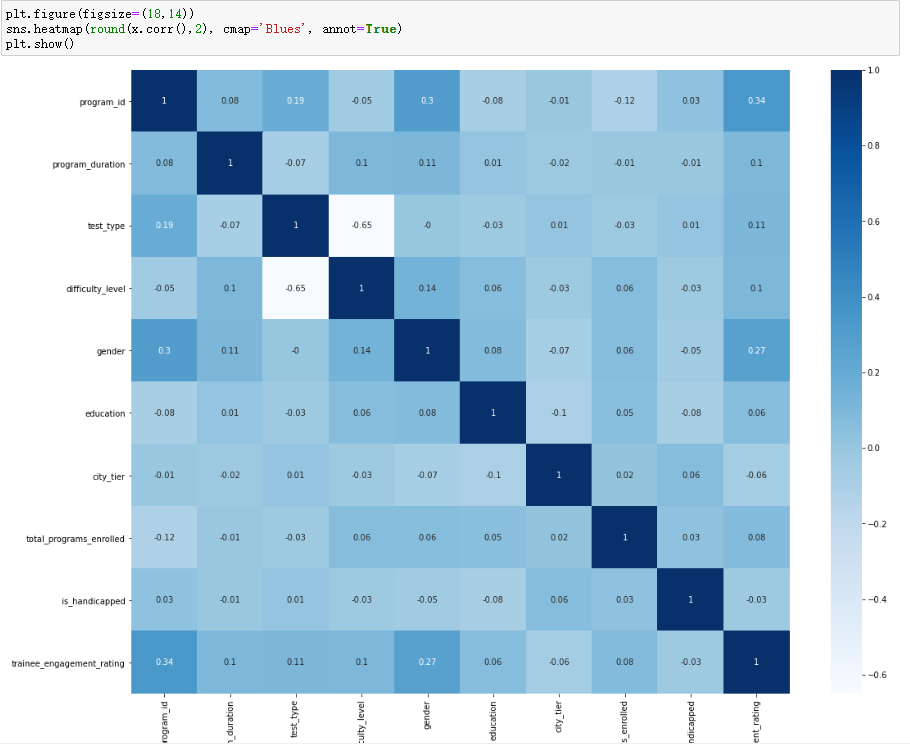


观察发现：

1、第11项特征“age”样本数缺失严重，但凭常识来讲，年龄这个特征十分重要，填充众数。

2、第0项特征“id\_num”与第4项特征“test\_id”和第7项特征“trainee\_id ”基本一致，故将第4项特征“test\_id”和第7项“trainee\_id ”去除，不利用该数据。

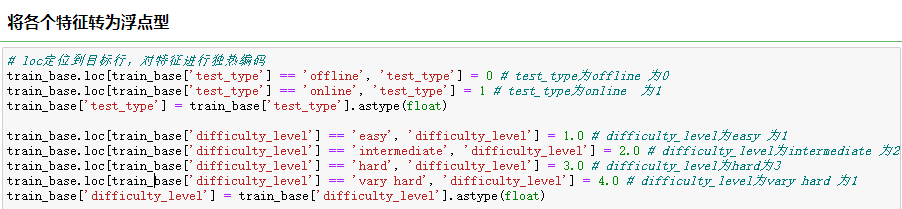
3、第3项特征“program\_id ”中包含第1项特征“program\_type”，故讲第1项“program\_type”去除，不利用该数据。



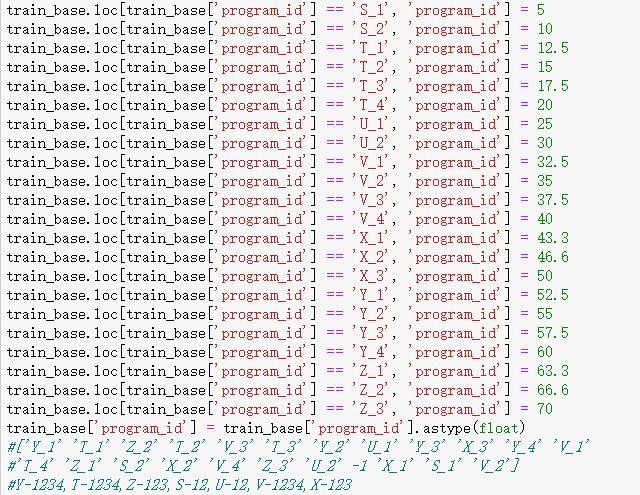
如上热力图，没啥冗杂的数据了。

**三、对数据进行预处理**

1、数据类型转化

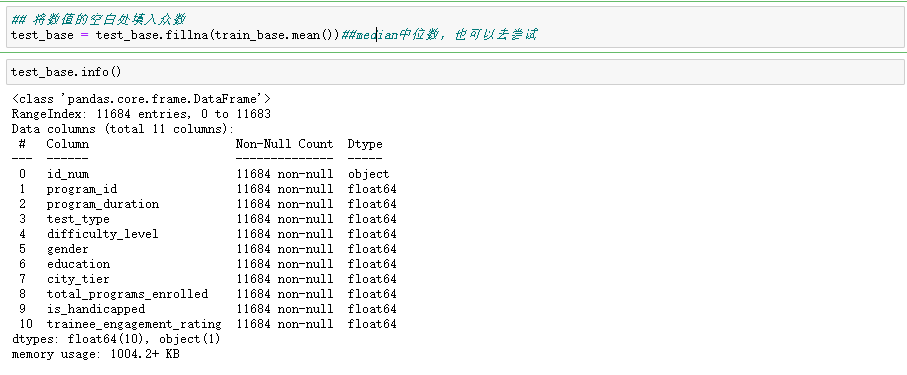


将各个object类型的数据，转化为float，方便观察，方便进行线性回归。

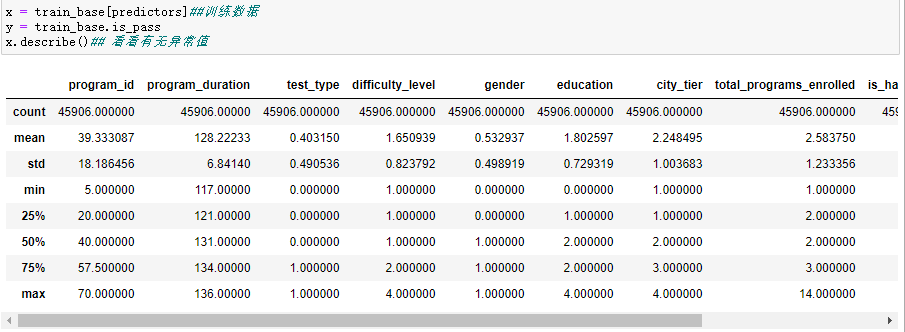


这边有个想法，本来每一个类型都是以1为间隔的，但是我觉得，如果同是一个字母类型比如T-1和T-2，他们内部的联系应该会紧凑一些，所以我以10为一个字母类型，数字平均分配空间，这样可能能加强他们内部的关系。

2、填补测试集中的缺失值。



3、查找有无异常值



观察有无异常值，例如太小太大的值；简单浏览发现，然而并没有。

特征工程还有许多方法：

比如对数转化，降维，或者建立新特征

我根据我们训练集数据，发现他们彼此独立性很高，也不好删除某个特征。稍微重复的特征都给我删了。

本来我还想结合时事 防止女拳，把性别也删了，但是想了想，性别对于这个具体环境：通过测试来讲，是十分重要的特征，不能删去。

由于我们是打榜练习，也不方便进行算法评估，所以只能先走一步算一步。

而且时间不多，主要是提交机会也有限，既要调参，又得试不同模型，最坑的是杯KNN坑了好多机会，还好师兄的服务器炸了，给了我很多机会hhhhhhh。

**四、尝试算法，模型优化，调整参数**

1、线性回归 LinearRegression



模仿继元师兄的代码，进行线性回归的训练，模拟。

最终线性回归的得分为0.70多，还是有很大空间的

线性回归具体模型和算法原理，另一个数学大佬师兄讲过了，这里我就不献丑技了。

这里我还发现一个问题，就是调整界限，就是>0.5=1那里，改变0.5到0.55有时可以得到更高的准确率，说明直接线性回归，误差还是挺大的，最好对我们的特征再进行处理。

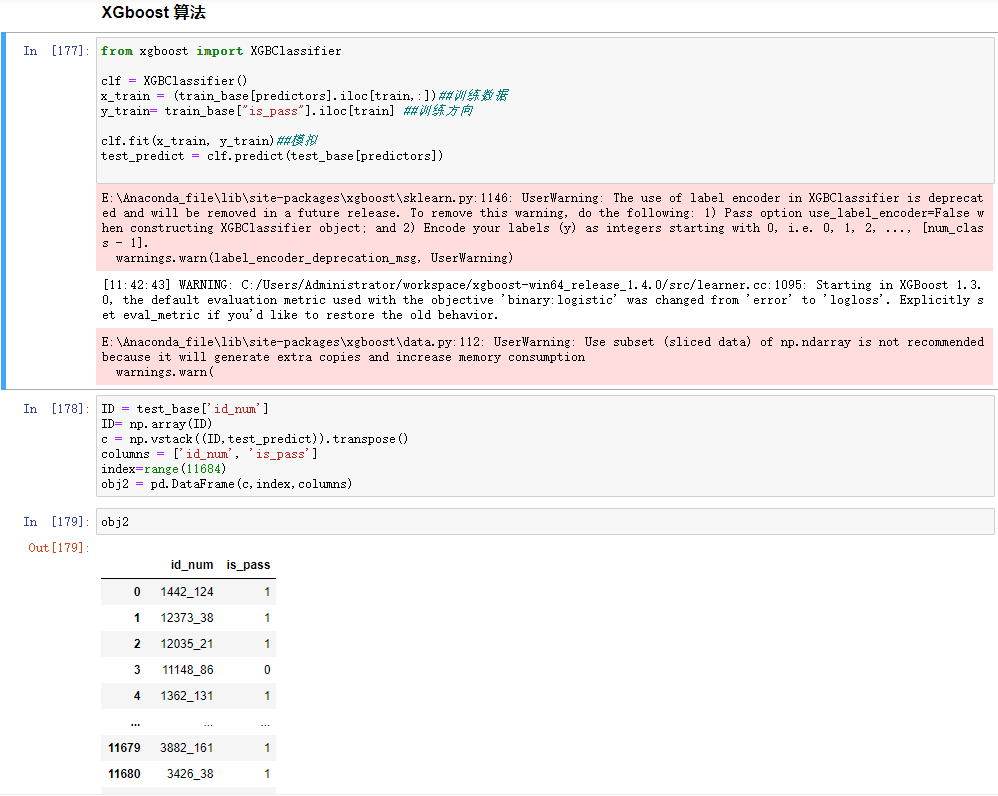
可能可以对数化

但是我现在水平有限，所学东西多且杂，暂时先放着，看看有没有更适合我们这个比较大数据集的算法。

2、XGboost 算法（借鉴了天池的实例）

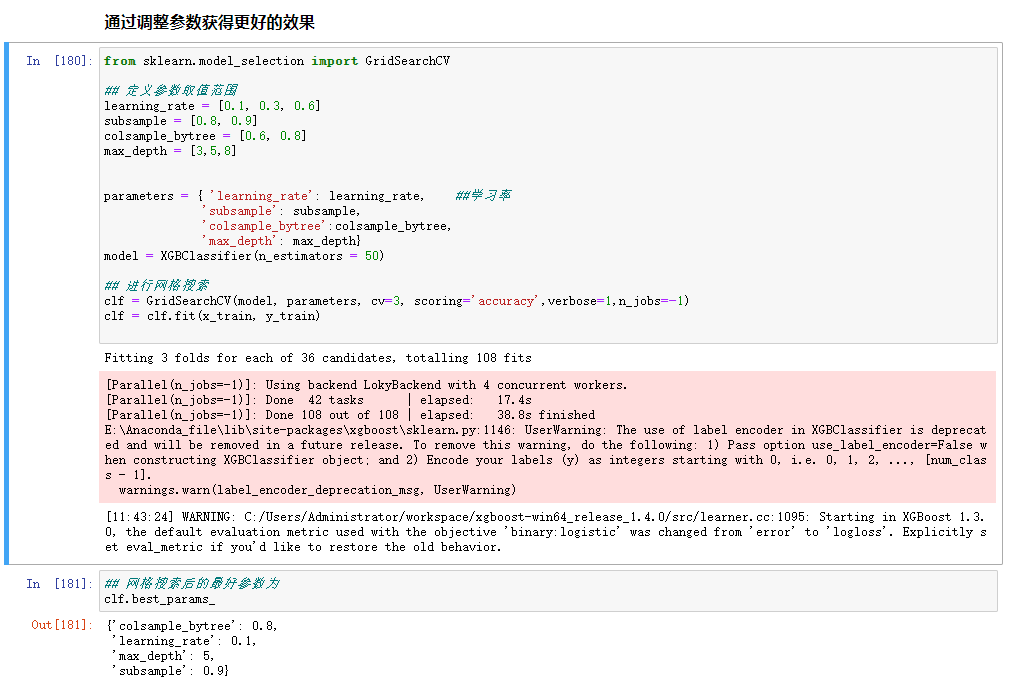
原文ID：<https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.20222472.J_3678908510.4.586a67c2npOV99&postId=170915>

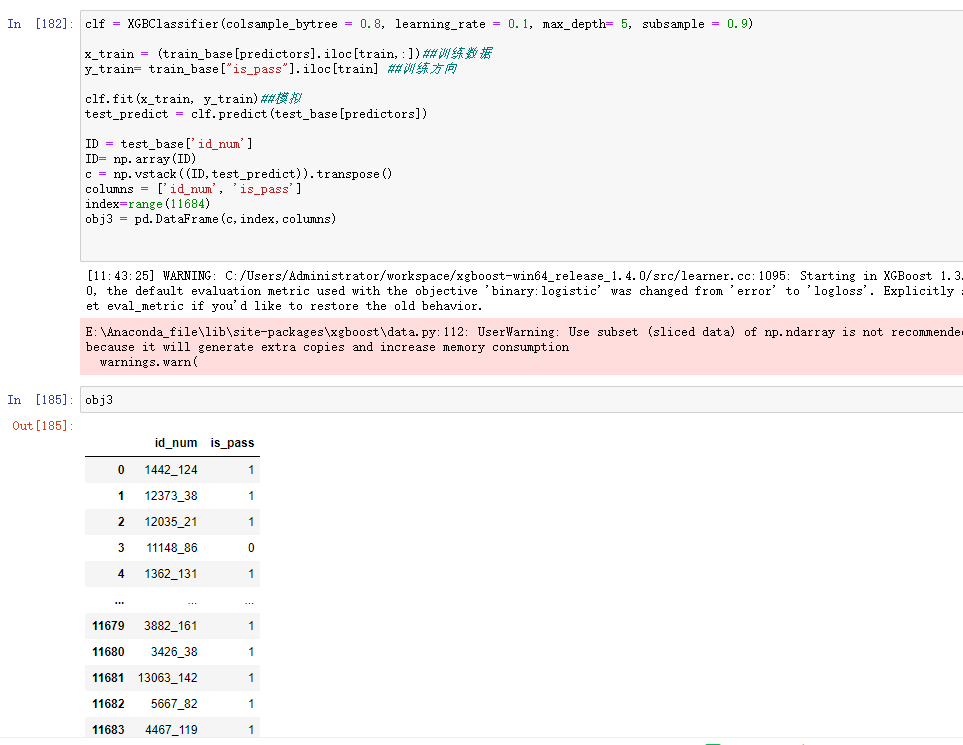
我在翻网上的资源时候，翻到了XGboost和LightGBM这两个东东，刚好有实例，跟我们的数据很相像，我决定换换接口，看看效果如何。

XGboost 它内部实现了梯度提升树(GBDT)模型。

未调参前XGBoost在A榜上大体得分是0.71964，比线性回归好很多

2.5 使用网格搜索法，调整参数。





经过调整，XGBoost在A榜上大体得分是0.72004，有一些提升

3、LightGBM 算法（借鉴了天池的实例）

流程与XGboost基本一致。







4、KNN算法

KNN算法是这里头最捞的了，浪费我不少提交机会QAQ。

本来我以为是K值取得不好，发现怎么取都是弟中弟，可能是数据的差过低，导致距离不明显，KNN算法就捞了。



**五、导出成果**



**六、不足之处**

1、未能掌握XGboost和LightBGM的核心原理，只是调用了接口（时间紧迫，需要一定时间学习）。

2、准确率没有上去，没有找到更好的处理数据方法。

3、寻找资源效率过低，前期在CSDN中浪费太多时间，最后还是去Kaggle和天池找具体实例模仿整个流程，才使效率提升起来。

4、由于没得找到标准答案，（我是猪，没有把train集分开当test用），所以算法模拟的时候只能靠榜单的分数自己脑补。。。结束了才想起来。

5、还有就是没有掌握好基础算法（决策树），会花时间补懂的。

6、时间有点急，学的东西太快了，可能以后会不扎实了。