****

****

**QG中期考核详细报告书**

**题    目      数据挖掘**

**学   院       自动化学院**

**专 业 大数据**

**年级班别 20级2班**

**学 号 3120001527**

**学生姓名 林桂鑫**

**2021年 4月16日**

目录

[数据清洗 3](#_Toc69512654)

[数据分析 3](#_Toc69512655)

[特征工程 5](#_Toc69512656)

[模型选择 6](#_Toc69512657)

[模型评估 6](#_Toc69512658)

# 数据清洗

先放代码：



1. 显示数据缺失

将数据输入后，我们得到一个panda中的Data Frame文件train\_data存放数据，然后用到missingno库中的matrix来显示缺失值，missingno中还有Bar Chart条形图用来展示无效数据的条形图，Heatmap展示变量缺失之间相关性，从-1到1，-1是指负相关，0指不相干，1指正相关。

二．通过缺失数据，将训练集中的不确定列（该列中我们所需的数据为空）去掉，或是将它们填充

1.用panda中的dropna将所需数据为non的列去掉。

subset括号里的索引中有Non值的给去掉，inplace=True是将源df给替换掉，axis=0/1(选择行或列)。

2.用.fillna(数据)进行填充

# 数据分析

先放代码：

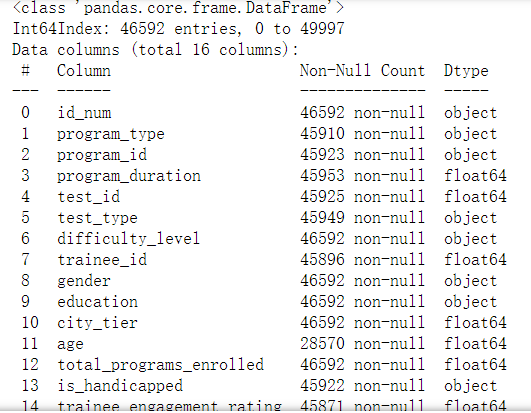


这是数据挖掘中比较重要的一部分，将数据类型转化为我们所需要的类型，一般要将string转为int或float。

1. 分析我们所需要的特征数据

这里我所需的特征数据是education , difficulty\_level , city\_tier , gender ，通过

train\_data.info() 的显示来看每列数据的类型



可以看到education , difficulty\_level , gender的类型是object，而不是数字类型，所以我们需要将它们转化为数据类型。

二． 转换数据(先定位，再赋值)

train\_data.loc[train\_data['education'] == 'Matriculation', 'education'] = 6

df.loc用于定位[[行]，[列]]，行是列索引education中值为Matriculation所在行，列为education，定位后可直接赋值。

# 特征工程

代码：



个人理解：特征工程即是将表面数据中没有的“隐藏数据”提取出来。

因为在考核中时间问题，没有写完特征age的处理—-—

1. 处理education

我的理解是，在大城市中，高学历的求职者数量比例较多，所以HR所注重学历的权重会下降，而注重学员的能力。在小城市中，高学历的求职者比例较小，HR会更看重学历，高学历等效于高能力。



所以利用city\_tier与education进行处理，得到一个新education，它代表相对学历（关于城市）。

1. 处理age

Age因为在训练集中缺失过多，如果不分青红皂白，直接取它的平均值来填补的话，可能模型表现更差了，所以我认为它不适合做此训练集的特征数据。当然，可以想别的办法来利用age，因为对于公司来说，受聘者的年龄也是十分重要的。所以我想制作一个适用于大多数训练集的age。想法：利用education（或其它数据），进行分类，比如学历为master中的平均年龄是38岁，则在education为master且age为空的格中放入38。这样分类来进行填空可能更好。

当然，我还有个想法，就是对age也做一个线性回归，通过其它的有age值的数据来推出无age值的空，仅供参考，我也不知道可不可以。

# 模型选择

部分代码：



利用sklearn.linear\_model库，建立一个线性回归类，通过fit函数将特征数据和目标数据进行线性回归类的训练，默认是用最小二乘法，然后用prediction函数对未知数据进行预测。

扩展：梯度下降法也能使用，需要用利用sklearn.preprocessing库中的StandardScaler（）进行归一化，.fit进行训练，.transform进行同样的归一化。在建立回归类要设定n\_iter为重复下降次数。

# 模型评估

代码：



应用：

通过sum可以从交叉验证中选择一个拟合最好的模型，



通过记录最优模型MaxNum，可以得到交叉验证中的最优模型alg1。

df.score(x,y)

(x,y) x是特征数据 y是结果(真实值)

计算R2



其中的y是真实值，f是预测值，\hat{y}是实际值的平均值

R2离1越近，说明模型的拟合程度越好，离0越近，说明拟合程度差。

可以当一个参考。