



# 廣東工業大學

## QG 中期考核详细报告书

题 目 中期考核  
学 院 计算机院  
专 业 计算机类  
年级班别 计算机类 6 班  
学 号 3120004892  
学生姓名 唐有成

2021 年 4 月 14 日

# 一、定义数据挖掘问题

题目描述：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。

题目分析：需要从给的数据集中的特征中筛选出一组对数据影响权重大的特征，并且使用这些特征构建模型预测学员是否通过。目标值只有通过和没通过，即 1 或 0，所以这是一个二分类问题。

# 二、数据查看与分析

查看数据集的前十五个数据，并且查看数据集的大小与特征数。

如图所示：

	id_num	program_type	program_id	program_duration	test_id	test_type	difficulty_level	trainee_id	gender	education	city_tier	age	total_programs
0	9389_150	Y	Y_1	136.0	150.0	offline	intermediate	9389.0	M	Matriculation	3.0	24.0	
1	16523_44	T	T_1	131.0	44.0	offline	easy	16523.0	F	High School Diploma	4.0	26.0	
2	13987_178	Z	Z_2	120.0	178.0	online	easy	13987.0	M	Matriculation	1.0	40.0	
3	13158_32	T	T_2	117.0	32.0	offline	easy	13158.0	F	Matriculation	3.0	NaN	
4	10591_84	V	V_3	131.0	84.0	offline	intermediate	10591.0	F	High School Diploma	1.0	42.0	
5	12531_23	T	T_3	134.0	23.0	offline	intermediate	12531.0	F	High School Diploma	1.0	29.0	
6	17874_144	Y	Y_2	120.0	144.0	online	easy	17874.0	M	Bachelors	2.0	48.0	
7	8129_61	U	U_1	134.0	NaN	online	easy	8129.0	M	Matriculation	2.0	45.0	
8	5652_57	U	U_1	134.0	57.0	offline	easy	5652.0	M	Matriculation	4.0	NaN	
9	17019_153	Y	Y_1	136.0	153.0	offline	hard	17019.0	M	Bachelors	3.0	28.0	
10	9932_80	V	V_3	131.0	80.0	offline	easy	9932.0	F	High School Diploma	3.0	NaN	
11	8543_31	T	T_3	134.0	31.0	online	easy	8543.0	F	High School Diploma	4.0	NaN	
12	15848_149	Y	Y_1	136.0	149.0	offline	intermediate	15848.0	M	High School Diploma	1.0	NaN	
13	423_126	Y	Y_3	135.0	126.0	offline	intermediate	423.0	M	Matriculation	3.0	NaN	
14	8404_27	T	T_3	134.0	27.0	online	easy	8404.0	M	Bachelors	2.0	NaN	

图 1 数据查看

	program_duration	test_id	trainee_id	city_tier	age	total_programs_enrolled	trainee_engagement_rating	is_pass
count	49323.000000	49273.000000	49259.000000	49298.000000	30619.000000	49306.000000	49226.000000	49998.000000
mean	128.229366	91.414345	9863.493128	2.249097	36.514256	2.583114	2.397818	0.696288
std	6.889967	51.307852	5716.490640	1.010896	9.045487	1.239399	1.326378	0.459864
min	117.000000	0.000000	1.000000	1.000000	17.000000	1.000000	1.000000	0.000000
25%	121.000000	45.000000	5051.500000	1.000000	28.000000	2.000000	1.000000	0.000000
50%	131.000000	91.000000	9665.000000	2.000000	40.000000	2.000000	2.000000	1.000000
75%	134.000000	135.000000	14618.000000	3.000000	45.000000	3.000000	4.000000	1.000000
max	136.000000	187.000000	20097.000000	4.000000	63.000000	14.000000	5.000000	1.000000

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 49998 entries, 0 to 49997
Data columns (total 16 columns):
#   Column                Non-Null Count  Dtype
---  -
0   id_num                 49998 non-null  object
1   program_type           49267 non-null  object
2   program_id             49299 non-null  object
3   program_duration       49323 non-null  float64
4   test_id                49273 non-null  float64
5   test_type             49296 non-null  object
6   difficulty_level       49295 non-null  object
7   trainee_id            49259 non-null  float64
8   gender                 49291 non-null  object
9   education              49296 non-null  object
10  city_tier              49298 non-null  float64
11  age                    30619 non-null  float64
12  total_programs_enrolled 49306 non-null  float64
13  is_handicapped          49280 non-null  object
14  trainee_engagement_rating 49226 non-null  float64
15  is_pass                49998 non-null  int64
dtypes: float64(7), int64(1), object(8)
```

图 2 数据查看

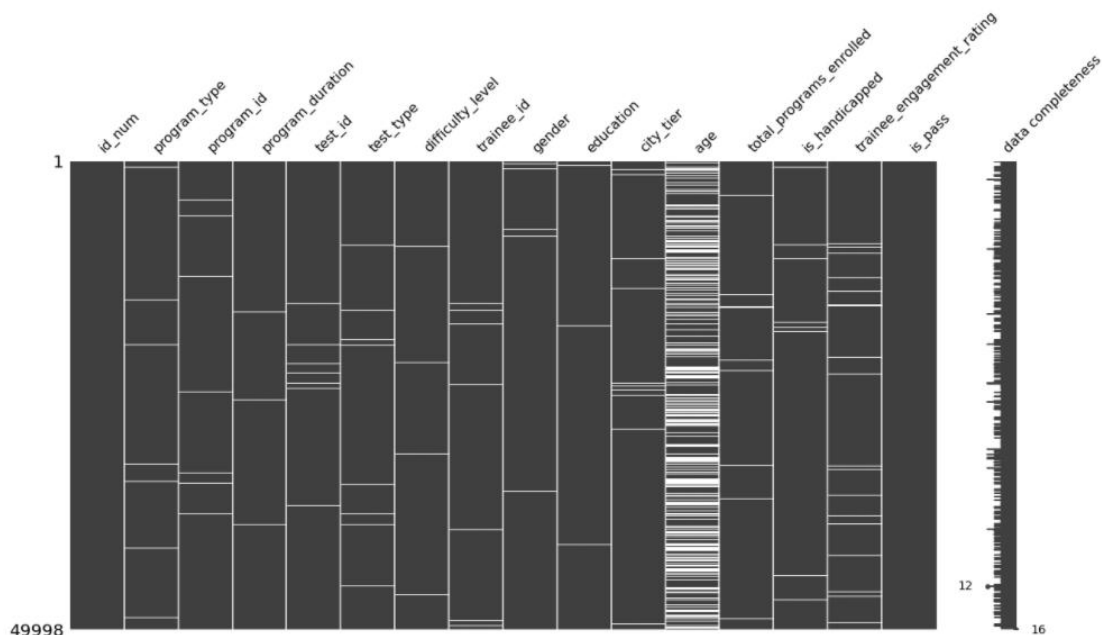


图 3 缺失数据查看

我们可以通过上面三张图发现，特征多，数据类型不全是数字型，需要对非数字型特征进行编码，并且各特征数据都有缺失，其中年龄特征的缺失最为严重，达到了接近 40%，其他特征数据缺失 1% 不到，如果我们对年龄这一特征进行平均值填充一定会对这一特征有很大

的偏差，构建模型也会因为填充的不确定性而导致对年龄特征平均值的过拟合，所以后续数据处理时选择去除年龄特征中值为空的数据。

### 三、数据预处理

由上面的数据查看与处分析，我们选择首先去除有空值的数据（之前走过弯路，用了平均值填充，后来明白缺失太多，均值也不一定准确，用均值填充也不正确）。处理后的数据如图所示：

	program_duration	test_id	trainee_id	city_tier	age	total_programs_enrolled	trainee_engagement_rating	is_pass
count	25427.000000	25427.000000	25427.000000	25427.000000	25427.000000	25427.000000	25427.000000	25427.000000
mean	128.308412	89.098478	10211.146144	2.235930	36.540449	2.519487	2.370433	0.696464
std	6.825941	52.085464	5914.145143	1.004349	9.037045	1.215236	1.333335	0.459793
min	117.000000	0.000000	1.000000	1.000000	17.000000	1.000000	1.000000	0.000000
25%	121.000000	43.000000	5135.500000	1.000000	28.000000	2.000000	1.000000	0.000000
50%	131.000000	87.000000	10091.000000	2.000000	40.000000	2.000000	2.000000	1.000000
75%	134.000000	133.000000	15387.000000	3.000000	45.000000	3.000000	4.000000	1.000000
max	136.000000	187.000000	20097.000000	4.000000	63.000000	14.000000	5.000000	1.000000

<class 'pandas.core.frame.DataFrame'>

Int64Index: 25427 entries, 0 to 49995

Data columns (total 16 columns):

#	Column	Non-Null Count	Dtype
0	id_num	25427 non-null	object
1	program_type	25427 non-null	object
2	program_id	25427 non-null	object
3	program_duration	25427 non-null	float64
4	test_id	25427 non-null	float64
5	test_type	25427 non-null	object
6	difficulty_level	25427 non-null	object
7	trainee_id	25427 non-null	float64
8	gender	25427 non-null	object
9	education	25427 non-null	object
10	city_tier	25427 non-null	float64
11	age	25427 non-null	float64
12	total_programs_enrolled	25427 non-null	float64
13	is_handicapped	25427 non-null	object
14	trainee_engagement_rating	25427 non-null	float64
15	is_pass	25427 non-null	int64

图 4 去除空值后数据查看

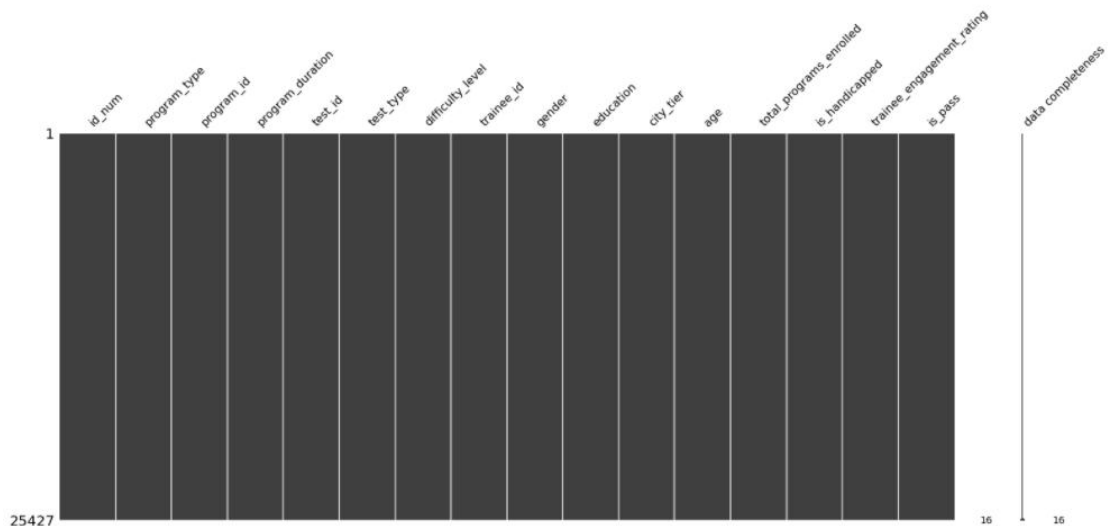


图 5 去除空值后数据查看

此时，我们得到了一份没有空值的数据集，虽然删去了许多数据，但是得到一份没有空值的数据集能够提高我们后面的模型模拟时的准确率。

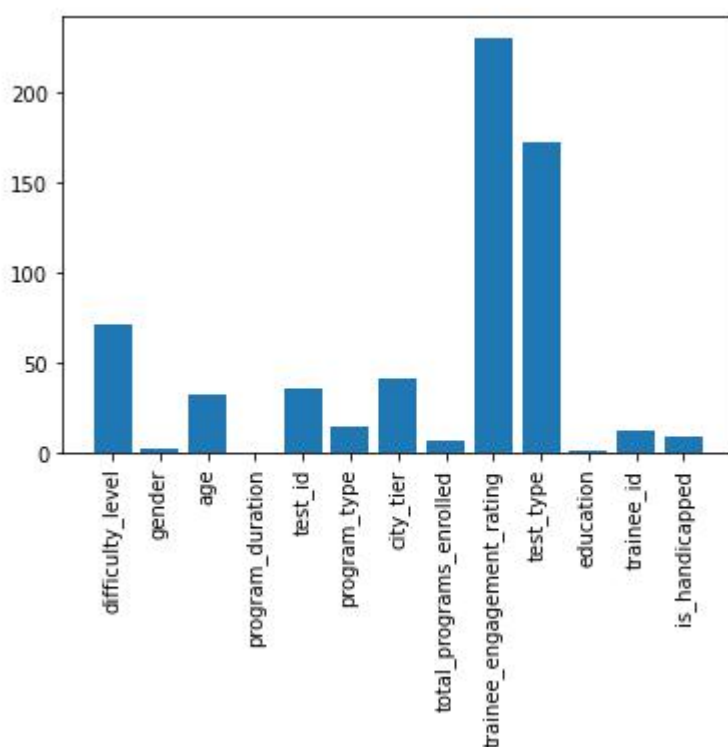
去除空值后，需要对非数值型数据进行编码，为后面的模型模拟做准备。首先我们对性别这一特征进行编码，而独热编码可以解决其他模型对一个特征的多个值标签编码带来的偏序，所以我们采用独热编码。同理我们对项目类型、考试类型、是否残疾这几个特征也采取独热编码。而考试难度与受教育程度应有一个大小的属性，所以我们采取标签编码，我们对考试难度分别用 0、1、2、3 对“easy”、“intermediate”、“hard”、“vary hard”编码，对受教育程度分别用 0、1、2、3、4 对 “No Qualification”、“High School Diploma”、“Matriculation”、“Bachelors”、“Masters”编码。所有数据编码完成后如图所示：

```
Data columns (total 16 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   id_num                                25427 non-null  object
1   program_type                          25427 non-null  uint8
2   program_id                            25427 non-null  object
3   program_duration                      25427 non-null  float64
4   test_id                              25427 non-null  float64
5   test_type                            25427 non-null  uint8
6   difficulty_level                      25427 non-null  object
7   trainee_id                           25427 non-null  float64
8   gender                               25427 non-null  uint8
9   education                            25427 non-null  object
10  city_tier                             25427 non-null  float64
11  age                                   25427 non-null  float64
12  total_programs_enrolled               25427 non-null  float64
13  is_handicapped                        25427 non-null  uint8
14  trainee_engagement_rating             25427 non-null  float64
```

图 6 编码后的数据信息

## 四、特征工程

当我们对所有特征编码完成后，需要筛选出对我们要预测的变量影响权重最大的几个特征。所以这里我们通过 sklearn 库 `SelectKBest` 函数进行单变量特征选取。最后再将结果可视化，如图所示：



这里我们首先选相关性最大的前 5 个特征，但是因为”age”与”test\_id”十分接近，我们也将其加进我们所需的特征，但是后续模型构建中，有了”age”特征准确率反而下降了，并且给出的测试集特征”age”下的值缺失数量大，补充后对模型预测也会有偏差，因此去除这一特征。

## 五、模型构建

在进行特征工程筛选出所需特征后，我们需要进行模型构建。这里我们对所有模型构建时都采用 K 折交叉验证这一方法将数据集分出训练集与测试集泛化模型，降低误差，同时降低过拟合的可能。

1、首先我们尝试简单的线性回归模型，通过设置阈值使它成为简单的二分类模型，这里我们阈值设为 0.5。准确率如图所示：

准确率为： 0.6836040429464743

2、因为这是一个二分类问题，所以接下来我们尝试逻辑回归模型，逻辑回归的基本思想与我们上面的线性回归设置阈值预测一样，但是逻辑回归替换了线性回归的回归函数，大幅减小离群值对回归函数的影响，使准确率提高。准确率如图所示：

准确率为： 0.7032681043311371

3、再接下来我们尝试使用随机森林模型进行分类，随机森林是基于决策树与 bagging 集成分类思想的一个模型，对缺失数据和非平衡数据非常稳健，所以这里我们采取这个模型进行尝试，准确率如图所示：

随机森林模型的准确率： 0.7177411530524536

4、因为加入逻辑回归分类器测试集准确率下降，所以最后我们选择 GBDT 和随机森林两个分类器取平均，提高预测的准确率。准确率如图所示：

0.6430565933849844

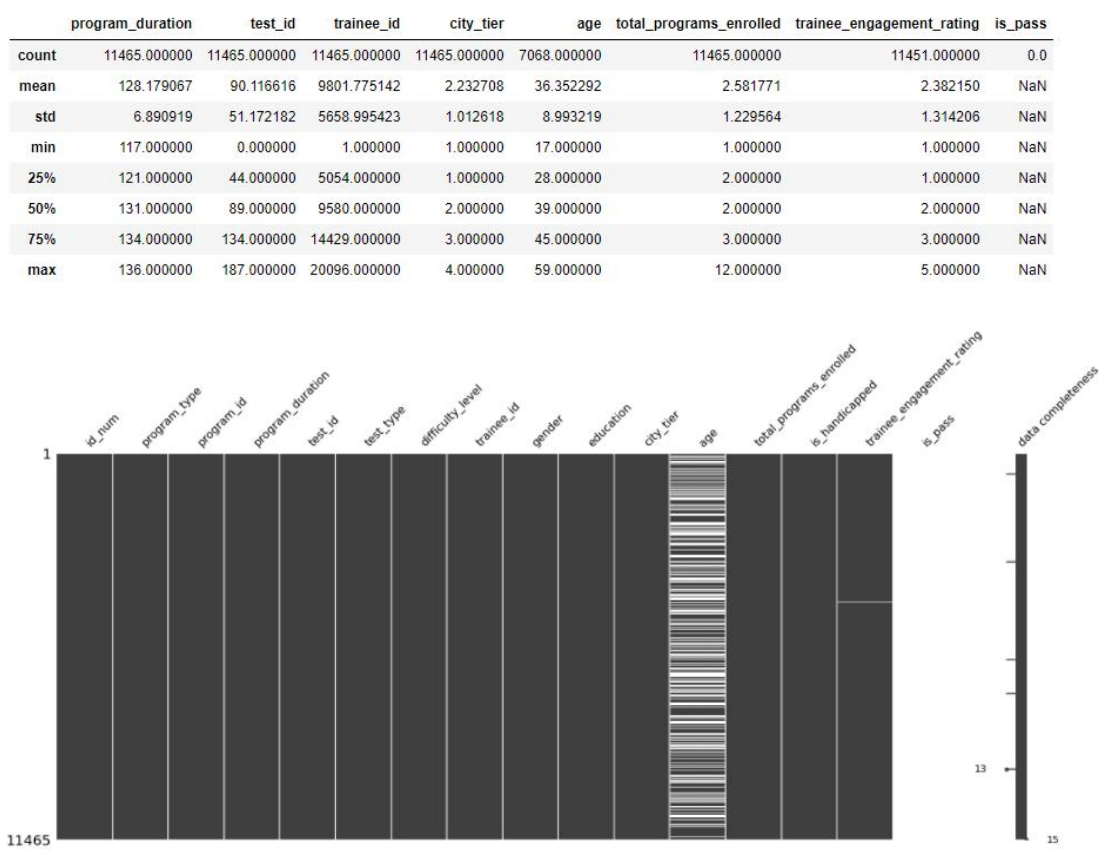
我们发现此时准确率反而大幅度下降了（相较于前面单个学习器），但是依照预期理论准确率应该上升，这个问题我询问了黄倬熙师兄，可是倬熙师兄让我自己研究研究，我又上 CSDN 上搜索，也没搜寻出有用信息，所以这里我给出我自己的一些思考见解：训练集使用 K 折交叉验证法准确率反而下降，其中有些许是分类器中各模型之间的关系问题，但是更大的原因是模型的泛化能力更强了，对训练集的预测反而失去了精度，模型对测试集等预测反而更加准确。为



了验证这一假设，接下来我们预测测试集。

## 六、预测测试集

1、导入测试集，查看数据信息与数据缺失。




2、数据处理

由上面数据查看发现，”age”特征数据有缺失，这里我们采取中位数填充，同时对非数字型数据采取与训练集一样的编码，完成后如图所示：

	program_duration	test_id	test_type	trainee_id	city_tier	age	total_programs_enrolled	trainee_engagement_rating	is_pass
count	11465.000000	11465.000000	11465.000000	11465.000000	11465.000000	11465.000000	11465.000000	11465.000000	0.0
mean	128.179067	90.116616	0.595726	9801.775142	2.232708	37.367728	2.581771	2.381683	NaN
std	6.890919	51.172182	0.490772	5658.995423	1.012618	7.177393	1.229564	1.313471	NaN
min	117.000000	0.000000	0.000000	1.000000	1.000000	17.000000	1.000000	1.000000	NaN
25%	121.000000	44.000000	0.000000	5054.000000	1.000000	30.000000	2.000000	1.000000	NaN
50%	131.000000	89.000000	1.000000	9580.000000	2.000000	39.000000	2.000000	2.000000	NaN
75%	134.000000	134.000000	1.000000	14429.000000	3.000000	43.000000	3.000000	3.000000	NaN
max	136.000000	187.000000	1.000000	20096.000000	4.000000	59.000000	12.000000	5.000000	NaN

### 3、预测训练集

已经将数据类型全都转换与训练出的模型一致，此时使用模型预测，准确率如图所示：



最高准确率 0.72272

## 七、后记

测试集的准确率如我们预期的一样比单一的分类器准确率提高了，证明了集成学习的优越性，但训练集中准确率下降的问题仍未解决，希望师兄能指点迷津。