数据分析及实践实验二实验报告

PB18061443 江昊霖

2023 年 5 月 3 日

1 实验目的

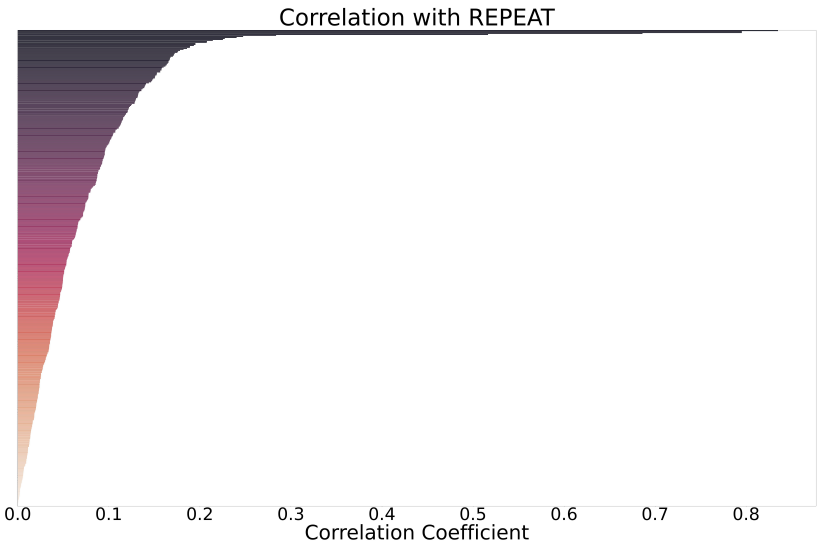
对 PISA2018 中的学生调查问卷数据集预测任务 REPEAT 列。

2 实验步骤

首先，使用 `pandas` 读取 CSV 文件，并查看数据集的基本信息：

```
<class 'pandas.core.frame.DataFrame'>
Index: 42176 entries, 0 to 42175
Columns: 486 entries, index to SOCONPA
dtypes: float64(471), int64(13), object(2)
memory usage: 156.7+ MB
```

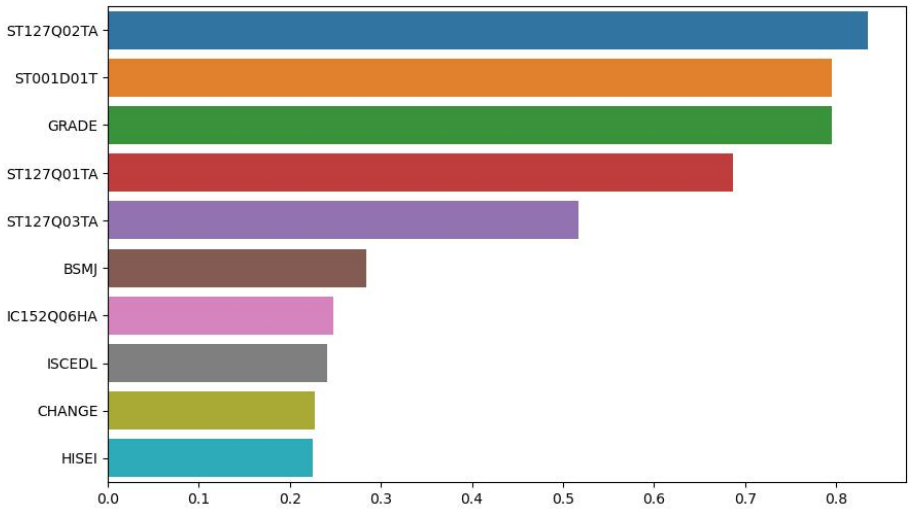
在查看了数据类型为 object 的两列之后，发现这些列是冗余的。因此，移除这两列，并针对所有列计算与 REPEAT 列的相关度。

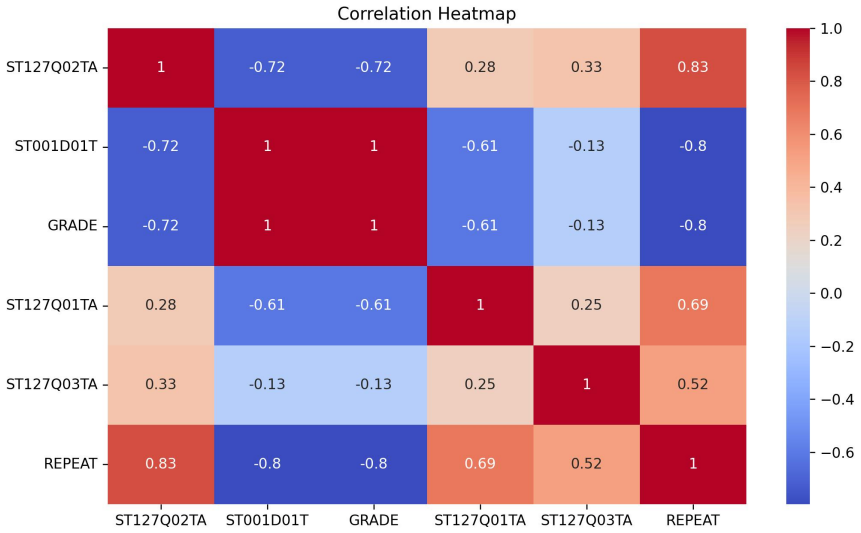


可以发现只有几列的相关系数较高，如下所示：

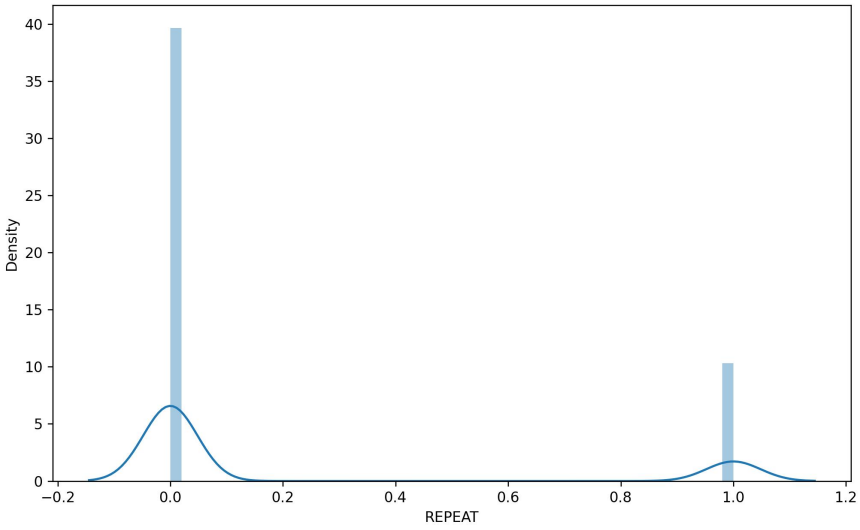
观察到 ST001D01T 和 GRADE 列的信息是相同的。因此，将 GRADE 列移除，并再次查看数据集的基本信息。

发现 ST127Q03TA 列存在较多的缺失值，决定移除该列。同时，还将包含 NaN 值的行移除。





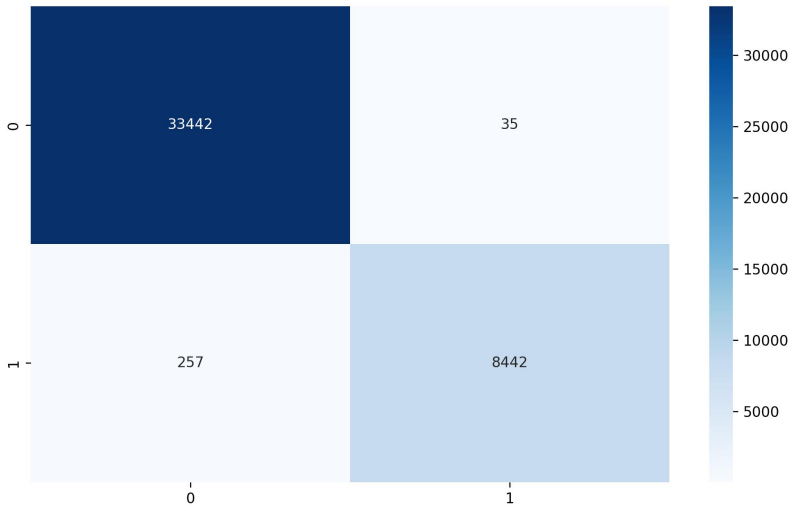

```
<class 'pandas.core.frame.DataFrame'>
Index: 42176 entries, 0 to 42175
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   ST127Q02TA  40543 non-null  float64
 1   ST127Q01TA  40055 non-null  float64
 2   ST127Q03TA  14385 non-null  float64
 3   ST001D01T   42176 non-null  float64
 4   REPEAT      42102 non-null  float64
dtypes: float64(5)
memory usage: 1.9 MB
```



从上图中，可以看出 REPEAT 列的数据分布只有 0 和 1 两种情况。因此，采用逻辑回归进行预测。

	precision	recall	f1-score	support
0	0.99	1.00	1.00	33477
1	1.00	0.97	0.98	8699
accuracy			0.99	42176
macro avg	0.99	0.98	0.99	42176
weighted avg	0.99	0.99	0.99	42176

经过以上步骤，成功地完成了数据预处理和逻辑回归预测。


33442
35
257
8442
0
1
0
1
30000
25000
20000
15000
10000
5000