

# PISA Data Analysis

姓名: 吴韬略  
学号: PB21051020

2023 年 5 月 4 日

## 摘要

## 1. 实验要求

- 实验要求
  - 给定一个数据集和预测目标，需要分析数据、统计以及抽取特征
  - 数据分析、统计，如：
    - 单个特征的分布
    - 统计缺失值
    - 特征间的相关性
    - 推测特征的含义
    - 异常样本
    - 数据抽样...
  - 特征抽取，如：
    - 特征的变换，如：str 转 int, 取log
    - 尝试组合特征
    - 特征子集选择
    - ...

(a)

### 实验要求

- 数据集-PISA2018 (简版版)
  - 本次数据针对PISA2018中的学生调查问卷数据集
    - 数据已经做了筛选，现包含语地区区的42176个学生，485个特征
- 
- 每个特征的具体含义可以参考codebook
  - Codebook是数据集每个特征的详细说明
  - 预测任务 REPEAT 列 (图例说明目标进行数据挖掘和特征工程)

(b)

图 1: 实验要求

## 2. 结果

自构建与 REPAET 强相关特征 SELF,EVERE,EVERE0-1,PARED,ENVIR。

- (a) 复读的学生一般 SELF 值较高, 即复读学生成就更高。
- (b) 复读学生的 EVERE 值集中在 2.5 左右, 未复读学生 EVERE 值集中在 -0.85 左右, 并且两者以 0 为严格界限。
- (c) EVERE0-1 与 REPEAT 二者同时为 1 或者 0, 完全等价。
- (d) PARED 值大很有可能会复读, 但是未复读学生的 PARED 取值无法预判。
- (e) ENVIR 值越大, 复读率越高。

# 目录

<b>1</b>	<b>方法</b>	<b>4</b>
1.1	预处理 . . . . .	4
1.1.1	数据类型转换 . . . . .	4
1.1.2	删除劣质列或者行 . . . . .	4
1.1.3	提取 REPEAT 列 . . . . .	4
1.1.4	去噪 . . . . .	4
1.1.5	空值填充 . . . . .	4
1.1.6	规范化 . . . . .	4
1.1.7	过滤式选择特征 . . . . .	4
1.1.8	对选择的特征再次预处理 . . . . .	4
1.1.9	处理异常值 . . . . .	5
1.1.10	数据预处理结果 . . . . .	5
1.2	特征构建 . . . . .	6
<b>2</b>	<b>数据分析</b>	<b>7</b>
2.1	相关系数矩阵 . . . . .	7
2.2	REPEAT 列数理统计 . . . . .	8
2.3	SELF 与 REPEAT . . . . .	8
2.4	EVERE, EVERE0-1 与 REPEAT . . . . .	9
2.5	PARED 与 REPEAT . . . . .	10
2.6	ENVIR 与 REPEAT . . . . .	11
<b>3</b>	<b>实验环境</b>	<b>12</b>
<b>4</b>	<b>结论</b>	<b>12</b>

# 1 方法

## 1.1 预处理

### 1.1.1 数据类型转换

使用 `sklearn.preprocessing.LabelEncoder` 类对字符型特征进行编码，即不同字符对应不同数字。

### 1.1.2 删除劣质列或者行

删除 REPEAT 为空值的行，空值大于 90% 的列，前两个索引列。

### 1.1.3 提取 REPEAT 列

用一个 numpy 数组储存 REPEAT 列后 drop 去除 REPEAT。

### 1.1.4 去噪

采用分箱去噪，每三个数据为一箱。

### 1.1.5 空值填充

采用中值填充空值。

### 1.1.6 规范化

通过 Z-score 规范化处理数据。

### 1.1.7 过滤式选择特征

通过各个特征与 REPEAT 列的皮尔森相关系数过滤特征，选择前 25 个特征。

### 1.1.8 对选择的特征再次预处理

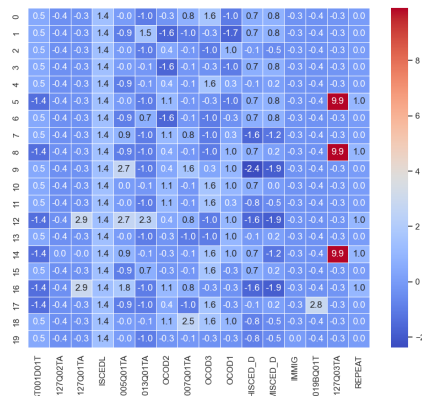
结合 CODEBOOK 一个个处理选择的特征，包括去除无效特征，把无效值转换为空值等。

## 1.1.9 处理异常值

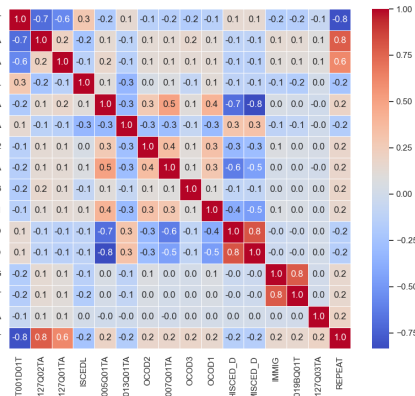
为了避免一次处理全部的异常值导致舍弃过多数据，处理异常值将在1.2特征构建后于每一个特征的分析中单独进行。方法：舍弃 $(\mu + 3\sigma, \mu + 3\sigma)$  之外的数据。

### 1.1.10 数据预处理结果

图2(a)以 heatmap 展示前 20 行的数据。



(a) 预处理结果



(b) 特征选择结果（协方差矩阵）

图 2: 数据预处理结果

## 1.2 特征构建

结合图2(b)中各特征的协方差和 CODEBOOK 中特征说明对特征进行归类并相加减：

1.  $SELF = ST001D01T + ISCEDL - OCOD3$

SELF：自己的学习成就。

2.  $EVERE = ST127Q01TA + ST127Q02TA + ST127Q03TA$

EVERE：之前的复读次数。

3.  $EVERE0 - 1 = \begin{cases} 1, EVERE > 0 \\ 0, EVERE \leq 0 \end{cases}$

EVERE0-1：由 EVERE 衍生的特征。

4.  $PARED = 5 - ST005Q01TA + 5 - ST007Q01TA + HISCED\_D + MISCED\_D$

PARED：父母的受教育程度。

5.  $ENVIR = -ST013Q01TA + OCOD2 + OCOD1 + IMMIG + ST019BQ01T$

ENVIR：生活环境，包括父母职业，移民身份，家中书籍，出生城市。

## 2 数据分析

### 2.1 相关系数矩阵

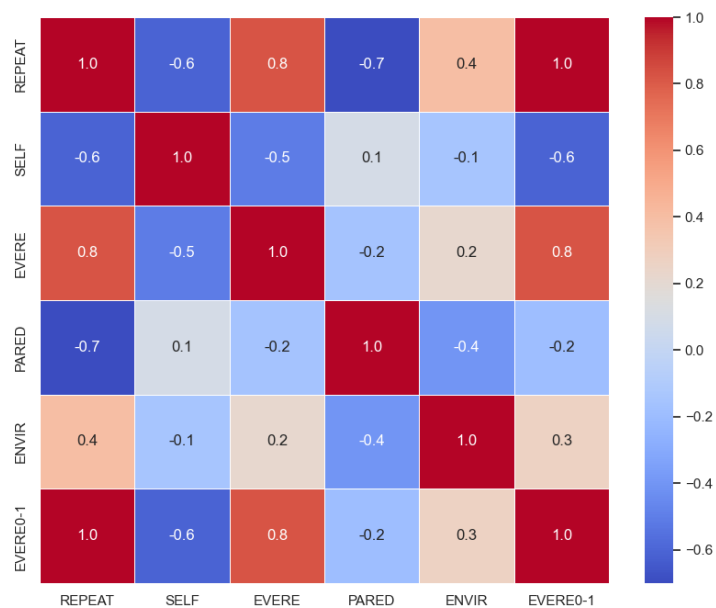
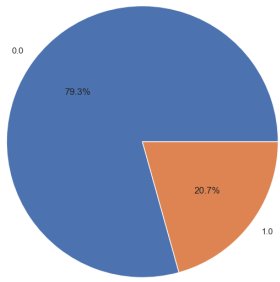


图 3: 自构建特征相关系数

REPEAT 与 SELF,EVERE,EVERE0-1,PARED,ENVIR 相关系数分别为-0.6,0.8,1,-0.7,0.4。构建的特征总体上具体较强代表性，并且各特征之间相关系数较低，说明两两之间独立性较强，无冗余特征。

## 2.2 REPEAT 列数理统计

count	mean	std	pie
42102	0.206617	0.404883	

REPEAT 人数只有未 REPEAT 人数的  $\frac{1}{4}$  左右，导致均值更加偏向于 0。

## 2.3 SELF 与 REPEAT

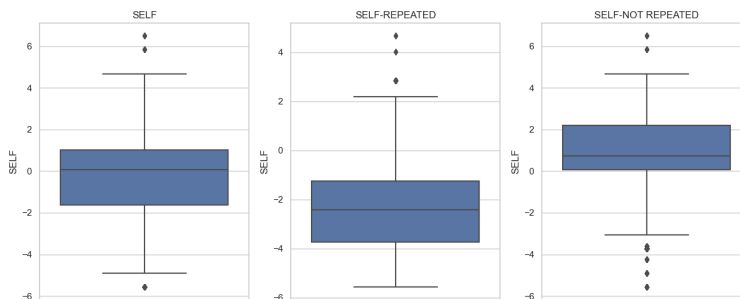


图 4: SELF-REPEAT

图4左，中，右分别为 SELF 特征总体分布，REPEAT=1 时 SELF 分布，REPEAT=0 时 SELF 分布的箱图。

三图对比明显说明复读的学生一般 SELF 值较高。这很好理解，复读学生付出了多倍努力，自然成就就会更高。或者说成就更高的学生读书时间往往更久，遇到的困难更多，有更大的概率复读。



## 2.4 EVERE, EVERE0-1 与 REPEAT

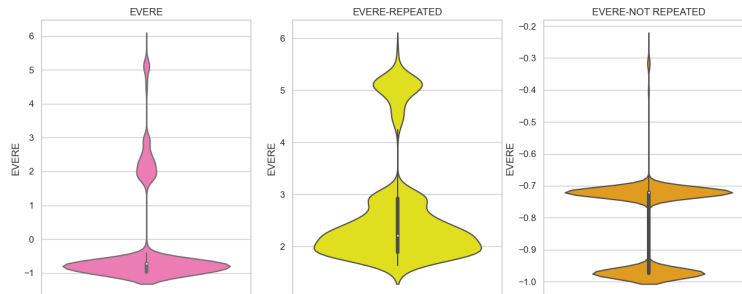


图 5: EVERE-REPEAT

图5左，中，右分别为 EVERE 特征总体分布，REPEAT=1 时 EVERE 条件分布，REPEAT=0 时 EVERE 条件分布的小提琴图。（由于之前用 Z-score 规范化处理数据，故 EVERE 可能为负数）

由于未复读学生是复读学生的四倍左右，故 EVERE 总体更加趋近于 NOT-REPEATED 分布，集中在小于 0 附近。

这是一个相当成功的特征，复读学生的 EVERE 值集中在 2.5 左右，未复读学生 EVERE 值集中在 -0.85 左右，并且两者以 0 为严格界限，可见即使是简单地使用 EVERE 是否大于 0 作为 REPEAT 值是否为 1 的分类判断条件，准确率也可以接近 100%。

由此我们可以构造出新特征 EVERE0-1，如1.2中所定义一样。

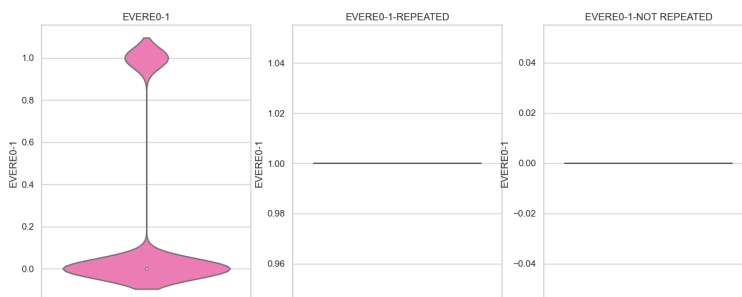


图 6: EVERE0-1-REPEAT

图6左，中，右分别为 EVERE0-1 特征总体分布，REPEAT=1 时 EVERE0-1

条件分布，REPEAT=0 时 EVERE0-1 条件分布的小提琴图。

果然，这是一个与 REPEAT 相关系数为 1 的特征，二者同时为 1 或者 0，完全等价！

## 2.5 PARED 与 REPEAT

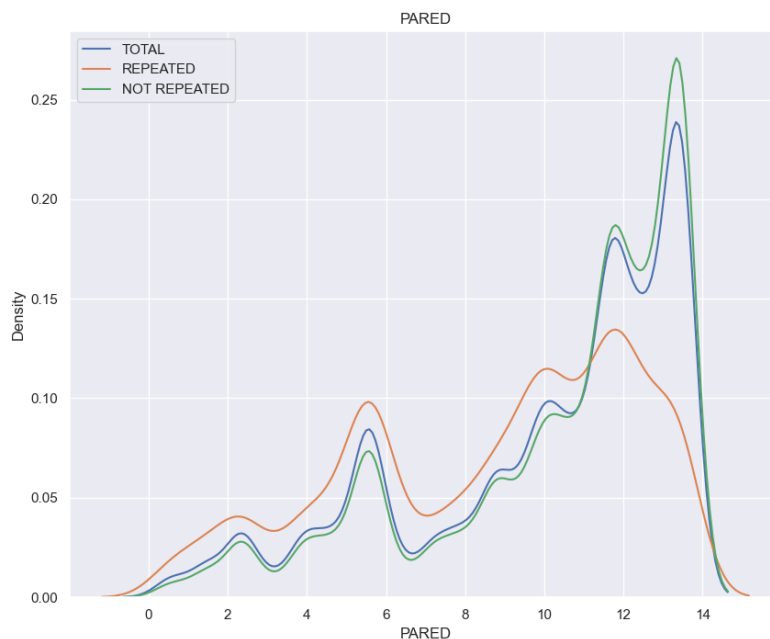


图 7: PARED-REPEAT

图7中，TOTAL 为 PARED 特征总体分布，REPEATED 为 REPEAT=1 时 PARED 条件分布，NOT REPEATED 为 REPEAT=0 时 PARED 条件分布的核密度图。

三线对比，复读学生的 PARED 值在高值处密度极大，未复读学生的 PARED 值则较为均匀，在中间和高值处密度都较大，说明 PARED 值大很有可能会复读，但是未复读学生的 PARED 取值无法预判。

## 2.6 ENVIR 与 REPEAT

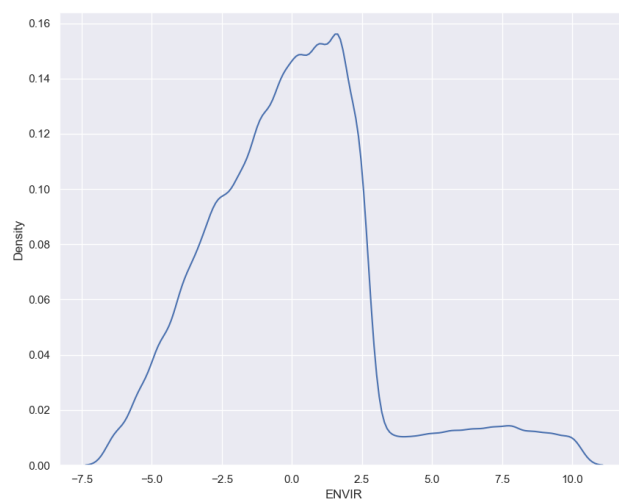


图 8: ENVIR

图8为 ENVIR 总体分布的核密度图，根据此图，确定 ENVIR 离散化分箱至范围  $[-7.5, 11)$ ，箱子长度为 2.5。

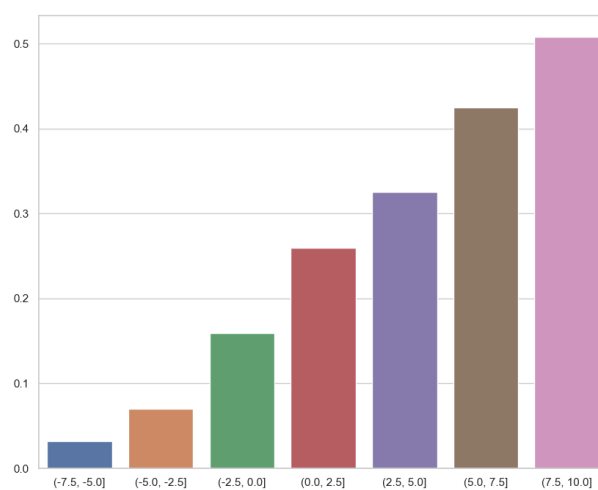


图 9: ENVIR-REPEAT

图9为不同 ENVIR 值区间的复读率条形图。

可见二者具有强相关性，ENVIR 值越大，复读率越高。古人曰：寒门出状元，诚如此言。

### 3 实验环境

```
1 import numpy as np
2 import pandas as pd
3 import sklearn.preprocessing as sp
4 from sklearn.impute import SimpleImputer
5 import matplotlib.pyplot as plt
6 from scipy.stats import zscore
7 import warnings
8 import seaborn as sns
```

图 10: 使用的库或者包

jupyter notebook; visual studio; anaconda

### 4 结论

1. 复读的学生一般 SELF 值较高, 即复读学生成就更高。
2. 复读学生的 EVERE 值集中在 2.5 左右, 未复读学生 EVERE 值集中在-0.85 左右, 并且两者以 0 为严格界限。
3. EVERE0-1 与 REPEAT 二者同时为 1 或者 0, 完全等价。
4. PARED 值大很有可能会复读, 但是未复读学生的 PARED 取值无法预判。
5. ENVIR 值越大, 复读率越高。

## 参考资料

- [1] QiLiu. Prof.qiliu' s slides on ad2023. <http://staff.ustc.edu.cn/qiliuql/-files/AD2023/2.2.pdf>. 2023-03-24.
- [2] QiLiu. Prof.qiliu' s slides on ad2023. <http://staff.ustc.edu.cn/qiliuql/-files/AD2023/2.3.pdf>. 2023-03-31.
- [3] QiLiu. Prof.qiliu' s slides on ad2023. <http://staff.ustc.edu.cn/qiliuql/-files/AD2023/2.4.pdf>. 2023-04-07.