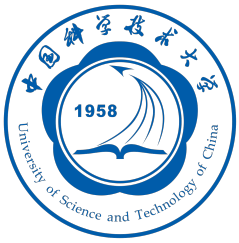


中国科学技术大学

University of Science and Technology of China

数据分析与实践 Lab3

Pisa2018 数据分析与特征处理

姓名：邱文韬

学号：JL22110003

专业：计算机科学与技术

学院：计算机科学与技术学院

2023 年 5 月 3 日

摘要

本文基于 PISA2018 数据集，首先对具有 485 列特征的原始数据集进行预处理，主要步骤为 1. 筛选特征：分区统计各个特征的缺失率，优先考虑缺失率低于 20% 的特征，其余特征采用插值法等进行填补 2. 异常值处理：根据大数定理判定异常数据，再对使用的特征使用填充替代或删除处理异常值，3. 数据类型分析：将字符串类型的特征删除或者转为数值类型。预处理结束后对数据进行分析统计，主要从相关性、数据分布特点以及异常样本等方面进行分析。然后进行特征抽取，通过协方差矩阵发现与 REPEAT 明显强相关的 5 个特征为 **ST001D01T：**Student International Grade (Derived)，**ST127Q01TA：** repeated At <ISCED 1>，**ST127Q02TA：** repeated At <ISCED 2>，**ST127Q03TA：** repeated At <ISCED 3>，**GRADE：** Grade compared to modal grade。通过特征组合得到新特征：**COUNT_1:** 学生家庭学习配套硬件情况，**INTERTIM：**学生使用互联网总时长，**Math~Music_sumtime**：学生分别在 Math 及 Music 等 9 门科目 (9 个特征) 上花费的总时间，**Major/Minor_SUM:** 在 sumtime 基础上继续总结“主副课”的学习时间进行分析，**PARIDX：**父母地位及自我期望综合指标，**FAMWEA:** 家庭经济条件指标。其中 **ST001D01T** 等 5 个特征与 **REPEAT** 表现出强相关，而 **PARIDX**、**COUNT_1** 与 **REPEAT** 与 **REPEAT** 存在关系，通过观察其余特征与 **REPEAT** 的密度分布等，比如其中 **FAMWEA** 的分布符合正态分布，我们认为其与 REPEAT 也存在一定的相关性。

关键字： PISA，数据分析，特征工程，相关性

目录

1 实验介绍 **3**

1.1 PISA 介绍 . . . 3

1.2 实验要求 . . . 4

2 数据预处理 **4**

2.1 认识数据 . . . 4

2.2 缺失值统计 . . . 5

2.3 特征异常比分析 . . . 7

2.4 特征数据类型分析 . . . 8

3 数据分析统计 **9**

3.1 相关性 . . . 9

3.2 学生国家、年龄及性别分布 . . . 11

4 特征抽取 **12**

4.1 COUNT_1: 学生家庭学习配套硬件情况 . . . 12

4.1.1 单特征分析 . . . 12

4.1.2 组合特征分析 . . . 14

4.2 IC 问卷信息：电子设备使用情况 . . . 14

4.3 IC 问卷组合特征：Math 等科目学习时间分析 . . . 17

4.4 Major/Minor_SUM：“主课/副课”学习总时间 . . . 19

5 连续变量的特征分析 **20**

5.1 连续变量的特征选择 . . . 20

5.2 特征分组 . . . 21

5.3 PARIDX：父母地位及自我期望综合指标 . . . 21

5.3.1 异常值处理及归一化 . . . 21

5.3.2 特征分析与组合 . . . 22

5.3.3 组合特征：PARIDX . . . 23

5.4 FAMWEA: 家庭经济条件指标 . . . 24

5.4.1 异常值处理及归一化 . . . 24

5.4.2 组合特征：FAMWEA . . . 26

6 总结 **27**

1 实验介绍

1.1 PISA 介绍

国际学生评估项目 (PISA)

对世界各地 **15 岁学生**进行的**三年一次调查**，评估他们获得了充分参与社会和经济生活所必需的**关键知识和技能的程度**。

覆盖范围广，被称为全球教育的“奥林匹克盛会”

问卷评估 + 认知项目测验

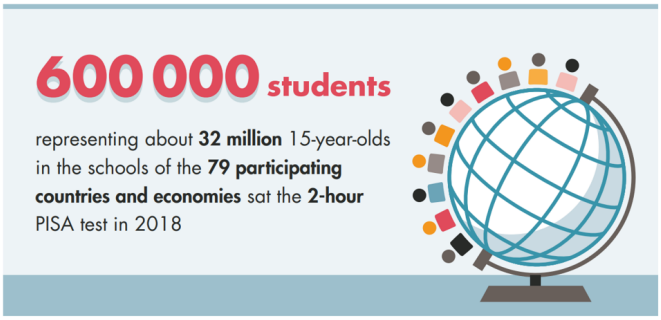


图 1: PISA 简介

具体覆盖范围

37 世界经合（OECD ）组织国家 + **42** 伙伴国家与经济体

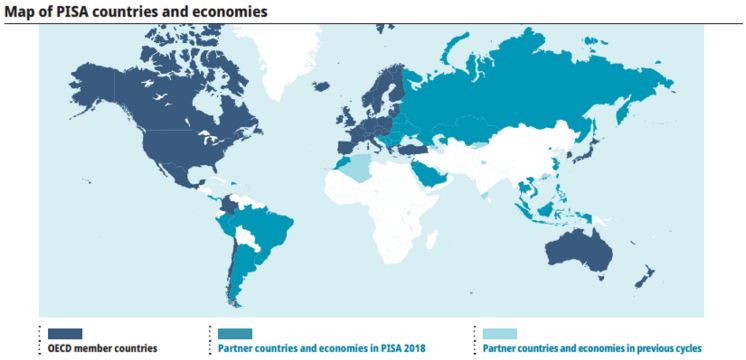


图 2: PISA 成员分布

OECD 学生能力国际评价项目

在国家层面上对学生总体做出推断，而不是评估学生个体的知识掌握情况

认知项目测验考察领域：科学、阅读、数学和金融素养

调查问卷：学生（含家长）+ 教师 + 学校

1.2 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
...

数据集

PISA2018（筛减版）
主要关注 PISA2018 中的学生调查问卷数据集
包含西语地区的 42176 个学生、485 个特征

预测任务

REPEAT 列-围绕预测目标进行数据分析和特征工程

2 数据预处理

2.1 认识数据

首先，因为我们的数据特征很多，有 485 个，并且这些特征基本都是一些语句、名词的缩写，或者是各个问卷中的问题的代码，我们很难直接从这些特征的名字直接了解这些特征描述的是什么，所以我在此首先将各个特征及其解释导入同一个表单中，便于查看以及分析这些特征的意义。

具体的工作如下：
首先从数据集 Codebook 中提取出各个特征的含义，作为 RawFeature.csv。接下来，从我

们的 lab3-data.csv 文件中提取出特征名称，作为 Lab.Feature.csv。因为总的特征实际上有 900 余个，而我们的实验数据集中需要分析的特征只有 485，所以不能直接简单的合并。所以在 jupyter 中根据特征名称将特征属性与之合并，结果写入 FeaAttr.csv 中。这样很方便查看每个特征及其含义，便于后续对特征进行分析。结果如下：

Feature and Attribute Table

Feature	Attribute
CNTRYID	Country Identifier
CNT	Country code 3-character
NatCen	National Centre 6-digit Code
STRATUM	Stratum ID 7-character (cnt + region ID + original stratum ID)
SUBNATIO	Adjudicated sub-region code 7-digit code (3-digit country code + region ID + stratum ID)
OECD	OECD country
ADMINMODE	Mode of Respondent
LANGTEST_COG	Language of Assessment
LANGTEST_PAQ	Language of Assessment (PAQ)
BOOKID	Form Identifier
ST001D01T	Student International Grade (Derived)
ST003D02T	Student (Standardized) Birth - Month
ST003D03T	Student (Standardized) Birth -Year
ST004D01T	Student (Standardized) Gender

图 3: 特征及其含义

2.2 缺失值统计

为了探索更多的特征，在此首先统计所有特征的缺失值分布。由于特征较多，如果在同一个图表中绘制，效果不佳，所以对特征进行分区统计。分区行号与特征如下表：

表 1: 特征缺失值分区统计

行号	特征内容
0:11	Country Id 等信息
11:71	Student 问卷
71:131	ICT familiarity 问卷
131:191	Educational career 问卷
191:251	Well-being 问卷
251:311	Financial literacy 问卷
311:371	Parent 问卷
371:486	学生与家庭相关信息

如图4，可以看出每个问卷的数据都有较多数据缺失，其中 ST 和 IC 问卷的数据最完

整，而其中 PA 问卷的缺失值是最多的。

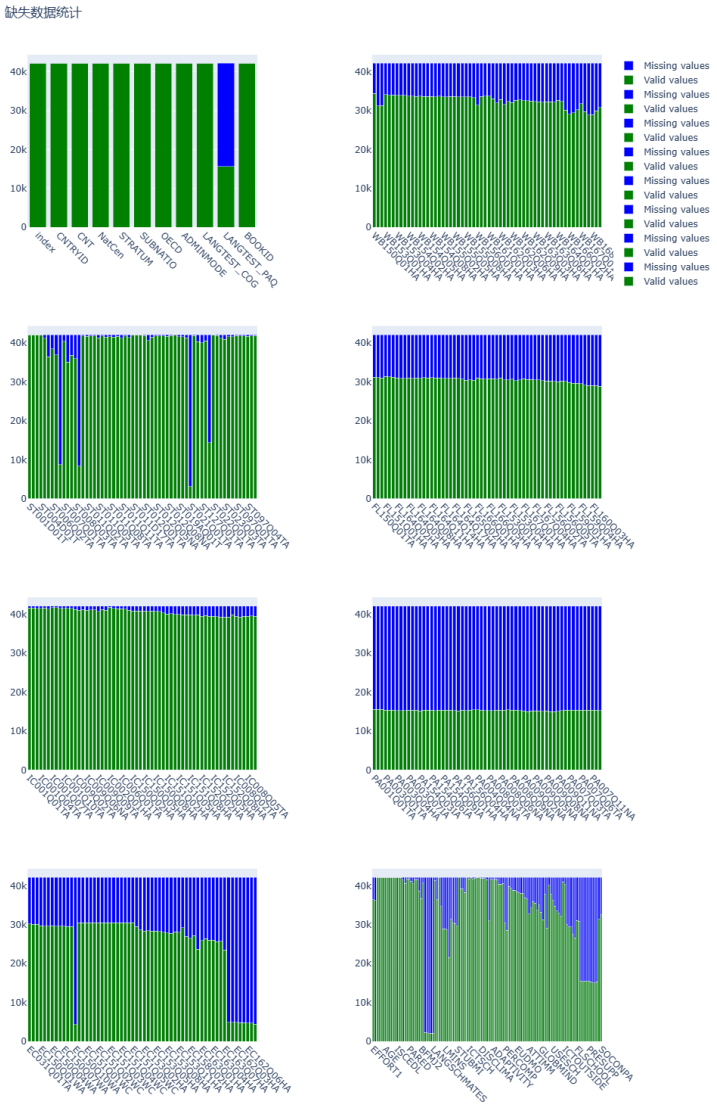


图 4: 所有特征缺失值统计

进一步查看每个特征的缺失比例，可以看出许多数据都有着较高的缺失率，其中如 LANGMOTHER、LANGFATHER 等 5 个特征缺失率达到了 95% 以上，而 Parent 问卷的所有特征缺失率都达到了 60% 以上，在此选择将缺失率达到 20% 以上的特征删去，并且这些缺失的特征主要是 EC、PA 问卷中的内容，还有一些关于 Parent 态度、支持的问题，相对其它特征重要性低一些，并且缺失率过高无法通过有效的填充达到还原分布特点的目的。

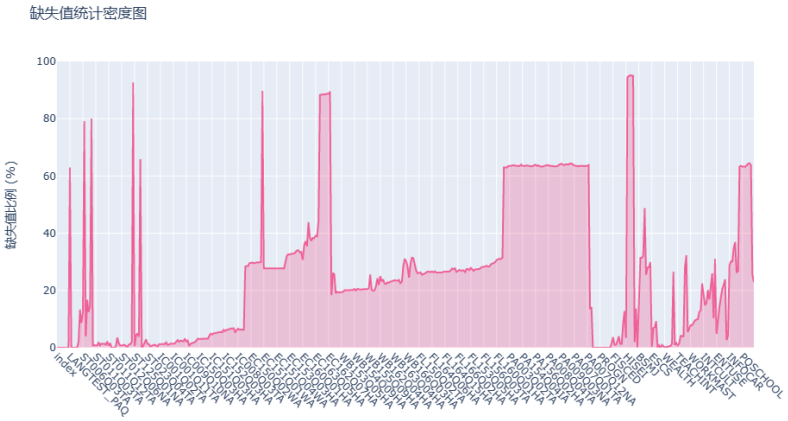


图 5: 特征缺失率

由于特征较多，所以在此先排除缺失率高于 20% 的特征 [1]，这部分特征缺失数据较多，即使填充也无法很好的还原其分布规律，所以在后续分析中优先考虑完整度更高的特征。排除缺失率高于 20% 的特征后，缺失率最高为 19% 左右，剩余特征数量为 207 列，密度图如下：后续的工作也优先分析这些较完整的特征与 REPEAT 的关系。

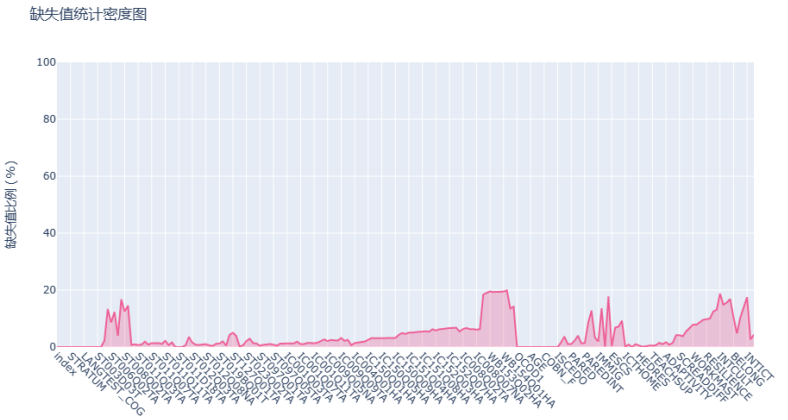


图 6: 排除高于 20% 缺失率后的特征缺失率分布

2.3 特征异常比分析

根据大数定理 [2] 通过判定超出 $(mean \pm 3 * std)$ 范围的数据为异常数据，如图7，可以看到有部分数据的异常比例在 5% 以上，但是其中有的数据为离散变量，并且大部分数据通过这个标准未检测出异常值，所以暂时不处理这部分异常值，在后续特征分析中如果用到某

些数据再进行具体分析。

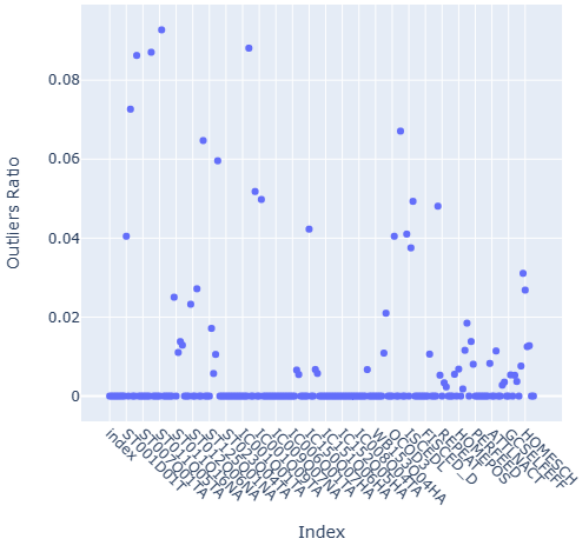


图 7: 特征异常比例分布

2.4 特征数据类型分析

观察数据各个特征的数据类型，如图8，发现绝大部分都是 float 类型，少数为 int，仅有两个属性是 object 类型的，这两个特征的信息与其它特征有冗余，并且是字符串，所以选择直接删去，其中，object 两列属性分别是：

1. CNT：Country code 3-character
2. STRATUM：Stratum ID 7-character (cnt + region ID + original stratum ID)

剩余数据都是数值类型，所以可以查看数据的方差分布，其中发现有一列 ADMINMODE 方差为 0, 查看 codebook 发现，ADMINMODE 的含义是：学生答题时使用的是电脑还是其它，说明数据中几乎所有人都是用电脑完成的答题，所以这一列特征对 REPEAT 明显没有作用，可以删去。

而其他的特征虽然方差比较低，但他们是一些对于学生家庭情况的问卷信息，所以在这里不直接删去，后续可以用于统计分析，帮助理解数据。

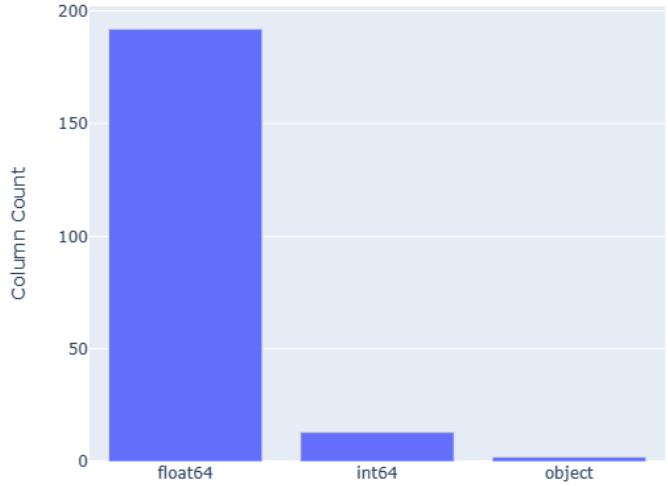


图 8: 特征数据类型

3 数据分析统计

3.1 相关性

由于初始特征较多，所以选择用散点图先观察每个特征与 REPEAT 的相关系数，如图9，可以看到，有部分特征呈现明显较强正负相关性，而其余大部分数据要么呈现不相关，要么集中在 ±0.2 之间：

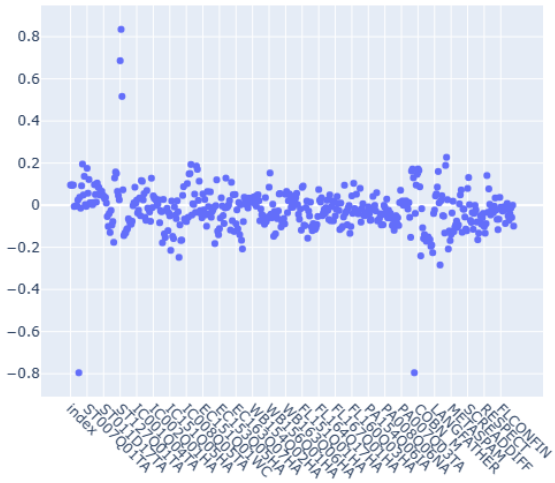


图 9: 各特征与 REPEAT 相关性

通过查看每个特征与 REPEAT 的相关性，可以看出 ST001D01T、ST127Q01TA-ST127Q03TA

以及 GRADE 与 REPEAT 有着强相关性。他们的含义分别是：

表 2: Attributes of Student Grades

Feature	**Attribute**
ST001D01T	Student International Grade (Derived)
ST127Q01TA	Have you ever repeated a <grade>? At <ISCED 1>
ST127Q02TA	Have you ever repeated a <grade>? At <ISCED 2>
ST127Q03TA	Have you ever repeated a <grade>? At <ISCED 3>
GRADE	Grade compared to modal grade in country

根据特征间相关系数的热力图10，可以看出 REPEAT 和这几个特征的确具有明显的强相关性

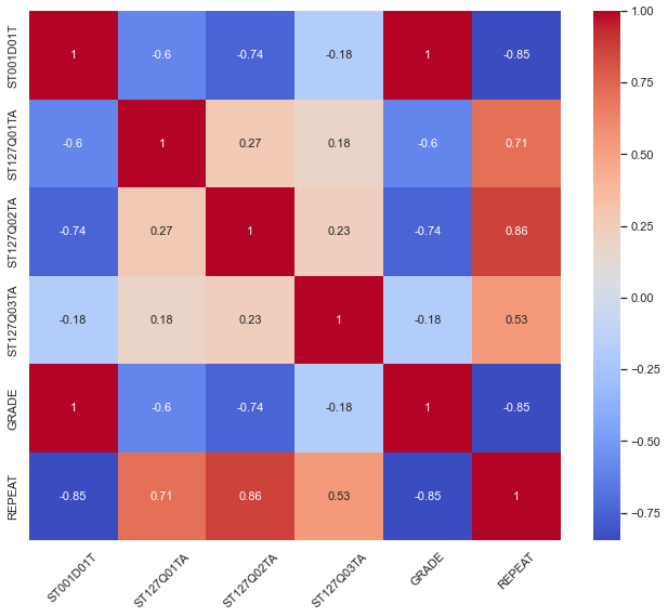


图 10: GRADE 等特征与 REPEAT 相关系数热力图

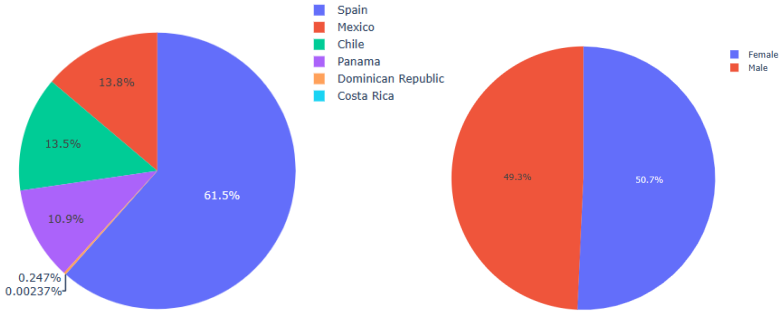
根据世界银行的 ISCED 分类指南 [3]，我推测特征 ST127Q01TA 是” 国际标准教育分类”(International Standard Classification of Education) 的缩写，1 2 3 等级分别代表：

1. 幼儿教育、学前教育、初等教育（小学）
2. 中等教育（初中、高中等）
3. 高等教育（大学、研究生等）

所以 ST127Q02TA 表现出最强的正相关，因为这个特征体现的就是问卷中学生是否在高中复读的问题。而如果在小学大学曾经复读过，那么说明学生的学习能力可能会弱一些，所以复读的可能性也更高一些，所以呈现正相关。而 ST001D01T 是学生的国际成绩水平，呈现较强的负相关，正符合成绩好的学生更少复读的直觉，GRADE 同理，它是和校内学生“模范”水平的比较，所以水平越高越不可能复读，所以呈现强负相关。

3.2 学生国家、年龄及性别分布

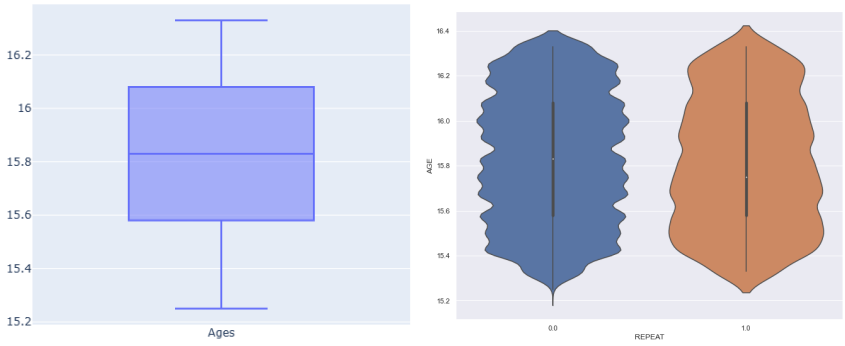
通过观察学生的国家、年龄和性别这 3 个特征，发现数据集中总共有 6 个国家，其中西班牙最多，超过了 60%，其次是墨西哥，智利等，这也和各个国家的发展情况相符合，更发达的国家参与测试的学生更多，性别分布则是女生稍多，但接近 1：1。



(a) 所属国家分布　　(b) 学生性别分布

图 11: 学生性别与所属国家分布

学生的年龄分布集中在 15-16 岁，并且与直觉不一致的是，是否复读与年龄并没有很大的联系，并不像预料中：复读的学生应该普遍比不复读的学生要更大一岁：



(a) 学生年龄分布　　(b) 学生年龄与 REPEAT 概率密度

图 12: 学生年龄与 REPEAT 分布

4 特征抽取

4.1 COUNT_1: 学生家庭学习配套硬件情况

4.1.1 单特征分析

从学生的家庭配套学习硬件情况可以看出他的学习环境情况, 选取特征 ST011Q01TA-ST011Q12NA 进行初步分析, 其中每个属性的值都是 1：yes 2：no

表 3: Description of Home Study Resources

Column Name	Description
ST011Q01TA	A desk to study at
ST011Q02TA	A room of your own
ST011Q03TA	A quiet place to study
ST011Q04TA	A computer you can use for school work
ST011Q05TA	Educational software
ST011Q06TA	A link to the Internet
ST011Q07TA	Classic literature (e.g. <Shakespeare>)
ST011Q08TA	Books of poetry
ST011Q09TA	Works of art (e.g. paintings)
ST011Q10TA	Books to help with your school work
ST011Q11TA	Technical reference books
ST011Q12TA	A dictionary

如图13，这 12 个特征缺失的数据都很少，相比于完整数据多于 40,000 的数据量，300 左右的缺失数据量可以忽略不计，所以直接删去这些缺失值。

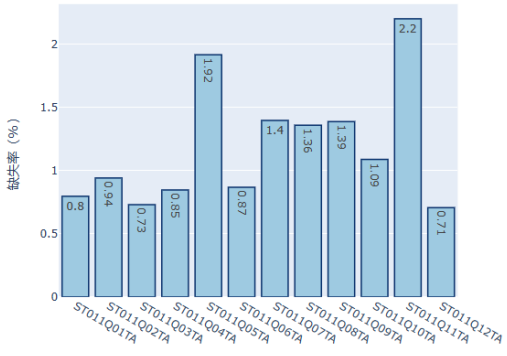


图 13: 学习配套硬件特征缺失率

如图14, 除了 Educational software、Classic literature、Books of poetry、Works of art、Technical reference books，其余 7 项占比均超过了 77%, 说明这些学生的家庭能够提供的学习环境与条件都不错，即使是拥有人数最少的 Educational software 也有 42.2% 的占比，说明大部分学生的学习不会受制于家庭能够提供的硬件条件。

图 14: 学习配套硬件拥有情况分布

如图15，其中 (ST011Q06TA:link to internet) 和 (ST011Q04TA: computer in home)，(ST011Q08TA:books for poetry) 和 (ST011Q07TA:classic literature) 有较强相关性，与他们的含义也是符合的，电脑与互联网连接，经典文学作品和诗歌书籍确实有较强的相关性。

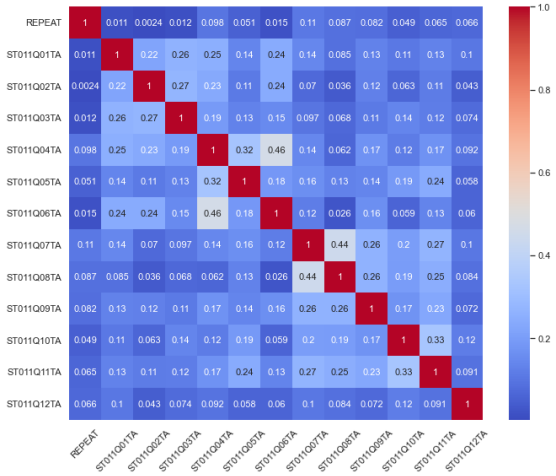


图 15: 学习配套硬件拥有情况相关系数热力图

4.1.2 组合特征分析

在这里，由于 1 表示学生拥有该设施，而 2 表示没有，所以为了量化学生的学习硬件支持水平 [4]，在这里统计每个学生这 12 个问题中 1 的个数表示该学生拥有的学习支持硬件水平，得到一个新的特征：COUNT_1。

如图16在没有复读的学生中，拥有配套硬件设施的数量较多的占大多数，并且不复读的学生中，有很多都是拥有 11-12 几乎所有硬件的学生，而复读的学生中，拥有的硬件数量显然更少，更偏向于中间。这也是符合直觉的，家庭条件更好，不会受到物质条件约束的学生学习阻力更小，因此复读的可能性也更低。

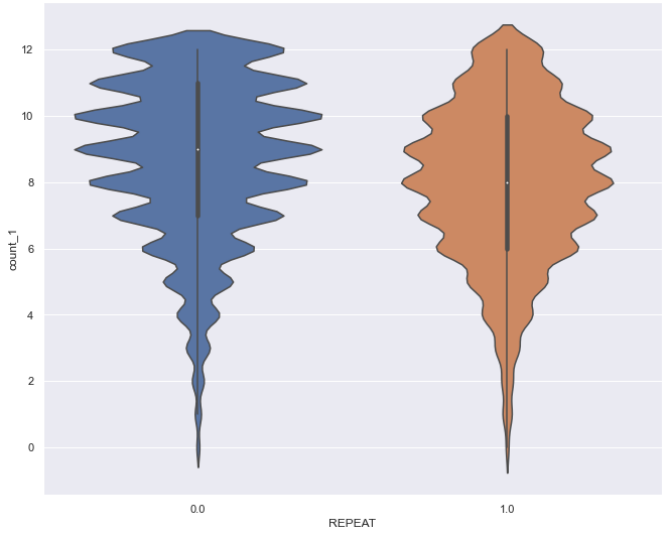


图 16: 组合特征 COUNT_1 与 REPEAT 密度分布

4.2 IC 问卷信息：电子设备使用情况

IC001Q01TA-IC009Q11NA、IC001Q10TA-IC009Q11NA 分别是是学生在家里、在学校能够使用的电子设备情况问卷信息，有三个取值，分别是：yes，yes but i don't use it，no。根据绘制的扇形图可以看出，大部分学生在家中或者在学校都有可用的各类电子设备，部分学生虽然可以用但选择不使用。

但其中也有部分设备超过半数的学生都没有，分别是：

1. 在家中：不能联网的智能手机、电子书阅读器 (e.g. Kindle)
2. 在学校：台式电脑、平板电脑 (e.g. iPad)、电子书阅读器 (e.g. Kindle)

这些设备相对于其它设备，会更昂贵一些，也能够找到一些其它的替代品，所以这可能是它们拥有率相对更低的原因。

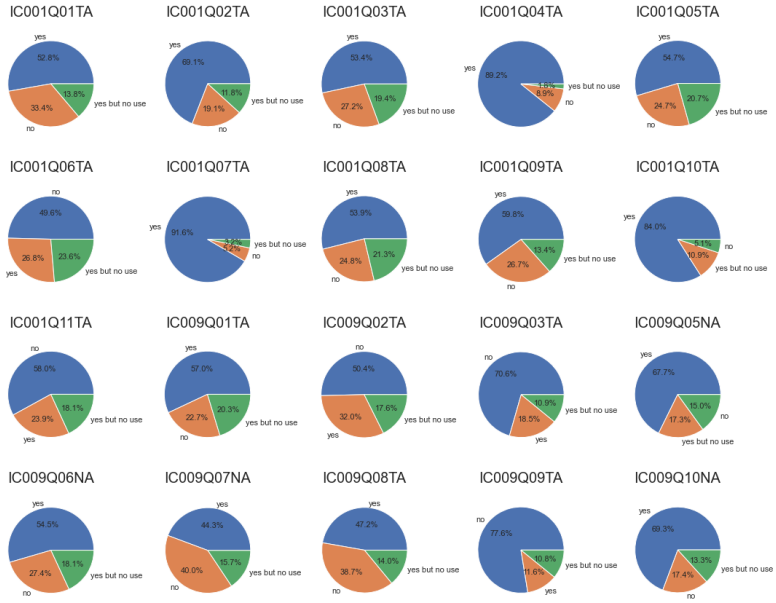
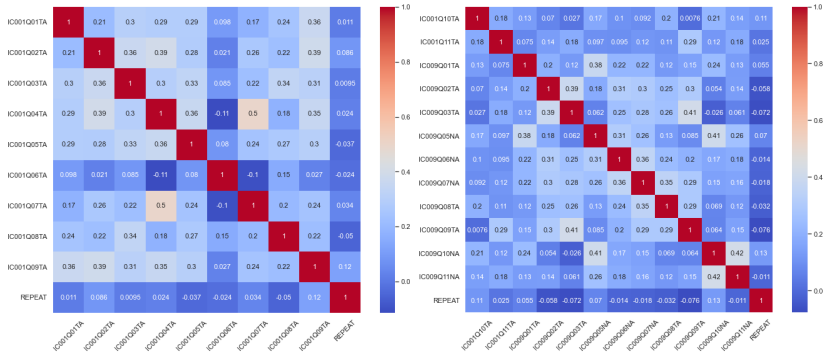


图 17: 学生电子设备使用情况分布

绘制上述特征与 REPEAT 的相关系数热力图，如图18，其中只有 IC001Q09TA，IC001Q10TA，IC009Q10NA 与 REPEAT 的相关系数相对更高，其他的相关系数都接近于 0，所以上述特征中只保留这三个特征。



(a) 校内使用情况　　(b) 校外使用情况

图 18: 学生电子设备使用情况与 REPEAT 相关系数分布

IC002Q01HA-IC007Q01TA 分别是学生第一次接触电子设备、互联网的时间，学生在工作日和周末在校内校外使用互联网的时间。如图19，其中，超过 60% 的学生在 4-9 岁第一次接触电子设备，而超过 70% 的学生在 7-12 岁第一次接触互联网。在校内时，超过 70% 的学生在工作日使用互联网的时间少于 1 小时每天。

而当在校外时，即使是在工作日，但学生使用互联网的时间相比于校内大幅上升，接近 70% 的学生每天使用的互联网的时间不少于 2 小时每天，有 23.1% 的学生每天使用互联网的时间超过了 6 小时每天。在周末时，这个比例进一步提高，接近 80% 的学生每天使用的互联网的时间不少于 2 小时每天，有 35.7% 的学生每天使用互联网的时间超过了 6 小时每天。

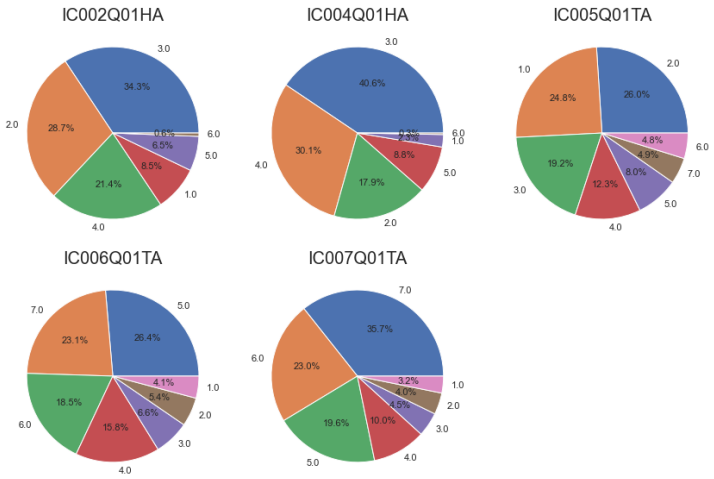
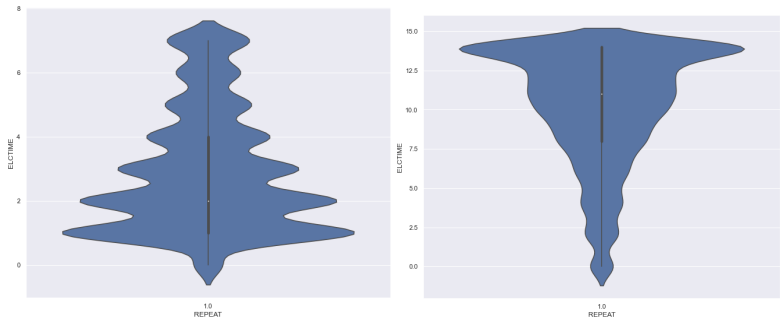


图 19: 学生第一次使用电子设备与上网时间分布

将学生使用互联网的时间相加 [5] 得到新特征 INTERTIME，观察这个新特征与 REPEAT 的关系，如图20。发现校内校外上网时长与 REPEAT 的关系并不一致。在校内，上网时长更短的学生复读概率反而更大，而在校外，上网时长更短的复读可能性变小了。我推测这是因为在校内使用互联网时更多的是用于学习，而在校外更多的是将互联网用于娱乐。



(a) 校内上网时长与 REPEAT　　(b) 校外上网时长与 REPEAT

图 20: 学生互联网使用时间与 REPEAT 密度分布

4.3 IC 问卷组合特征：Math 等科目学习时间分析

IC150Q01HA-IC152Q09HA 总共 27 个特征，它们分别代表的是在课内、课外以及校外，分别在 Math 等 9 个科目上花费的时间，具体如下表4:

表 4: IC Codes by Subject

Subject	Test ID
Test language lessons	IC150Q01HA, IC151Q01HA, IC152Q01HA
Mathematics	IC150Q02HA, IC151Q02HA, IC152Q02HA
Science	IC150Q03HA, IC151Q03HA, IC152Q03HA
Foreign language	IC150Q04HA, IC151Q04HA, IC152Q04HA
Social sciences	IC150Q05HA, IC151Q05HA, IC152Q05HA
Music	IC150Q06HA, IC151Q06HA, IC152Q06HA
Sports	IC150Q07HA, IC151Q07HA, IC152Q07HA
Performing arts	IC150Q08HA, IC151Q08HA, IC152Q08HA
Visual arts	IC150Q09HA, IC151Q09HA, IC152Q09HA

将这些特征按照科目进行分类，再将每个科目的时间度量相加，得到 9 个新特征。它们分别代表，学生在课内、课外以及校外，学习各个科目花费的总时长，如表5:

表 5: New Features by Subject

New Feature	Subject
Tll_sumtime	Test language lessons
Math_sumtime	Mathematics
Sci_sumtime	Science
Flan_sumtime	Foreign language
Ssci_sumtime	Social sciences
Music_sumtime	Music
Sport_sumtime	Sports
Parts_sumtime	Performing arts
Varts_sumtime	Visual arts

观察这 9 个特征的分布，对于 Math，Science，Language 等” 主课”，只有最多 14% 的学生每周学习 0 分钟或者不学习该科目，而 Musci，Arts 等 “副课” 则有 48.8%-62.9% 的学生根本不花时间学习该科目, 如图21:

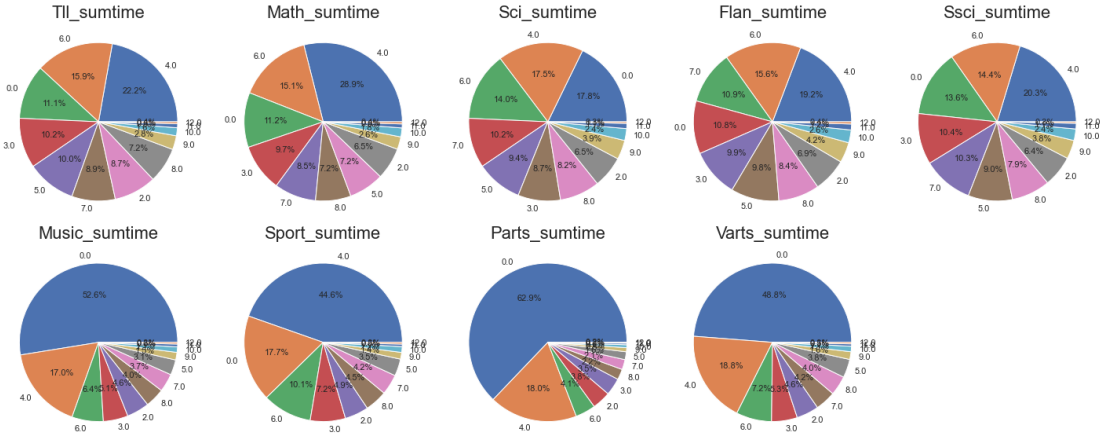


图 21: 各科目学习总时间分布

观察得到的 9 个新特征与 REPEAT 的相关性，如图22，其中 Music 花费时间与 REPAET 的相关性最高，其次是 Arts 等，也很容易理解，当学生花费更多时间在 Music 这类“副课时”，而 Math 这类“主课”的时间变少了，因此他们更可能会复读。

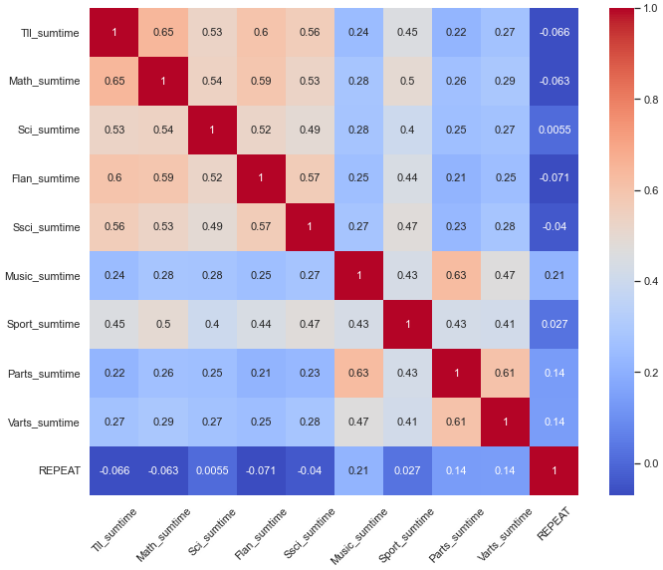


图 22: 科目学习总时间与 REPEAT 相关系数

因为这 9 个特征间的相关性也比较强，尤其是分别在 Math，Science，Language 等”主课”以及 Musci，Arts 等“副课”间，所以继续提取特征，得到两个新的“主课/副课”学习总时间：Major_SUM、Minor_SUM 特征，将在下一小节介绍。

4.4 Major/Minor_SUM：“主课/副课”学习总时间

根据前述，将“主课”、“副课”分别归类为：

1. “主课 (Major)”：
   - Test language lesson
   - Mathematic
   - Science
   - Foreign language
   - Social sciences
2. “副课 (Minor)”：
   - Music
   - Sports
   - Performing arts
   - Visual arts

分别观察“主课”总时间 (Major_SUM)、“副课”总时间 (Minor_SUM) 与 REPEAT 的密度分布，可以看出“副课”对 REPEAT 的影响要更明显一些，如图23,24：

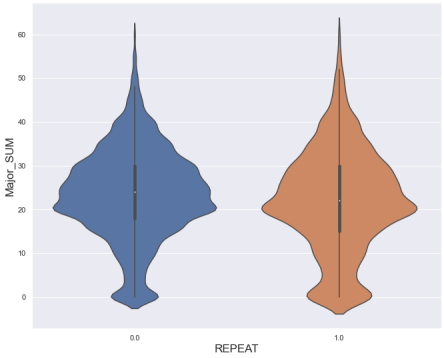


图 23: Major_SUM 与 REPEAT 密度分布

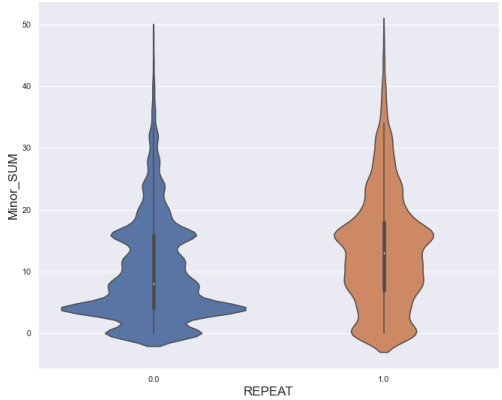


图 24: Minor_SUM 与 REPEAT 密度分布

5 连续变量的特征分析

5.1 连续变量的特征选择

以下是主要的一些连续变量：缺失率高于 20% 的特征 (STUBMI, CHANGE, SCCHANGE，MMINS，LMINS，SMINS，TMINS) 暂不考虑

表 6: Description of selected variables

Variable	Description
PAREDINT	Index highest parental education (international years of schooling scale)
BMMJ1	ISEI of mother
BFMJ2	ISEI of father
HISEI	Index highest parental occupational status
BSMJ	Students expected occupational status (SEI)
~~MMINS~~	Learning time (minutes per week) - Mathematics
~~LMINS~~	Learning time (minutes per week) - Test language
~~SMINS~~	Learning time (minutes per week) - Science
~~TMINS~~	Learning time (minutes per week) - in total
~~SCCHANGE~~	Number of school changes
~~CHANGE~~	Number of changes in educational biography (Sum)
~~STUBMI~~	Body mass index of student
ESCS	Index of economic, social and cultural status
UNDREM	Meta-cognition: understanding and remembering
METASUM	Meta-cognition: summarising
METASPAM	Meta-cognition: assess credibility
~~ICTHOME~~	ICT available at home
~~ICTSCH~~	ICT available at school
HOMEPOS	Home possessions (WLE)
CULTPOSS	Cultural possessions at home (WLE)
HEDRES	Home educational resources (WLE)
~~WEALTH~~	Family wealth (WLE)

如图25，ICTSCH 和 ICTHOME 与 REPEAT 的相关性过低，接近于 0，所以舍弃这两个特征, 而 HOMEPOS 和 WEALTH 的相关性达到了 0.91, 说明这两个属性冗余了，所以只选取完整度更高的 HOMEPOS。

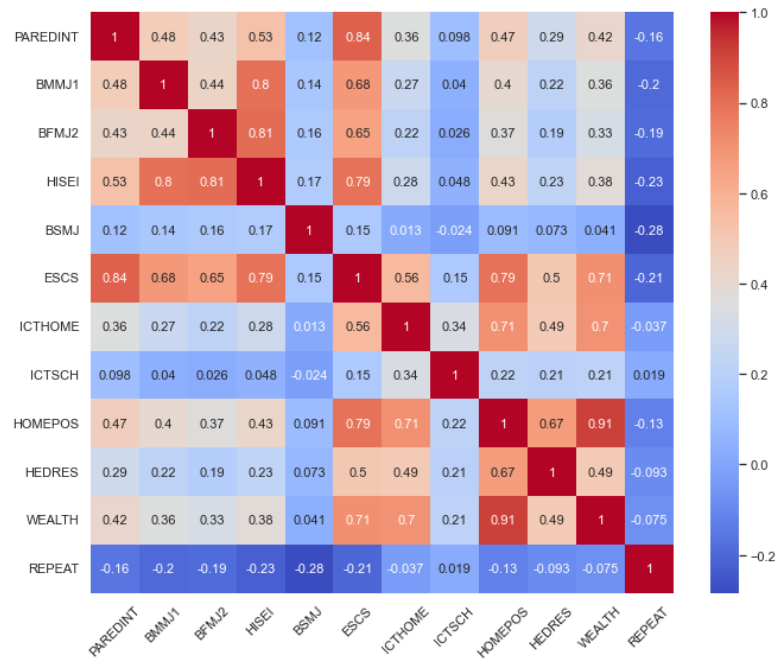


图 25: 主要连续变量特征与 REPEAT 相关系数

5.2 特征分组

手动对特征进行分组，按照特征含义及其取值范围，便于对数据进行缺失值填充以及归一化等操作。

1. 父母地位及自我期望 (PARIDX):
   - PAREDINT
   - BMMJ1
   - BFMJ2
   - HISEI
   - BSMJ
2. 家庭经济条件 (FAMWEA):
   - ESCS
   - HOMEPOS
   - HEDRES

5.3 PARIDX：父母地位及自我期望综合指标

5.3.1 异常值处理及归一化

因为这几个衡量学长家长教育及职业社会地位 [6] 的度量不一致，所以要进行归一化，但归一化前要进行异常值处理，因为归一化对异常值很敏感。如图26, PAREDINT 和 BSMJ 存在异常值，通过判断 (mean±3 * std) 找出异常比例分别占 1.0% 和 0.3% 左右，因为异常

数据并不是很多，所以选择将异常值也视作缺失值，用平均值填充 [7]。

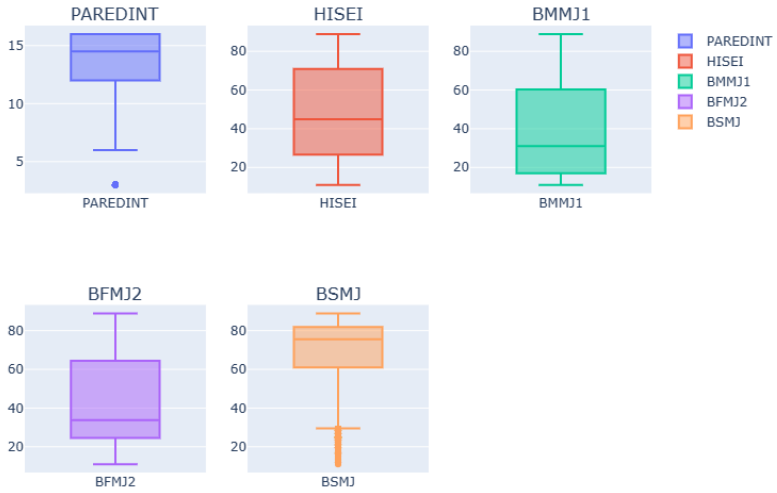


图 26: PAREDINT 等特征数据分布

处理完异常值后使用 sklearn 进行归一化至区间 [0,1]。

5.3.2 特征分析与组合

观察处理完后这 5 个特征的分布特点，如图27：

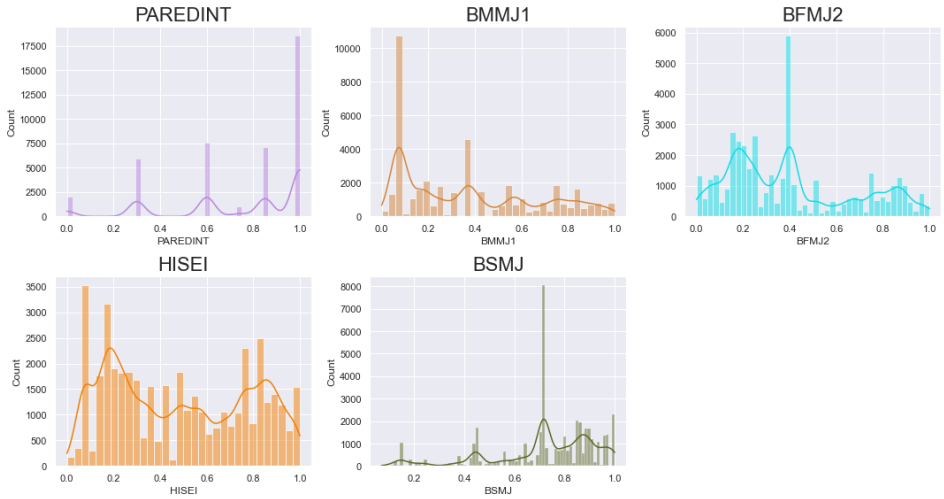


图 27: 处理后 PAREDINT 等特征数据分布

可以看到每列数据的分布没有什么明显规律，不是正态分布。取对数后，数据的分布更密集，但仍不呈现正态分布或其他分布规律，如图28：

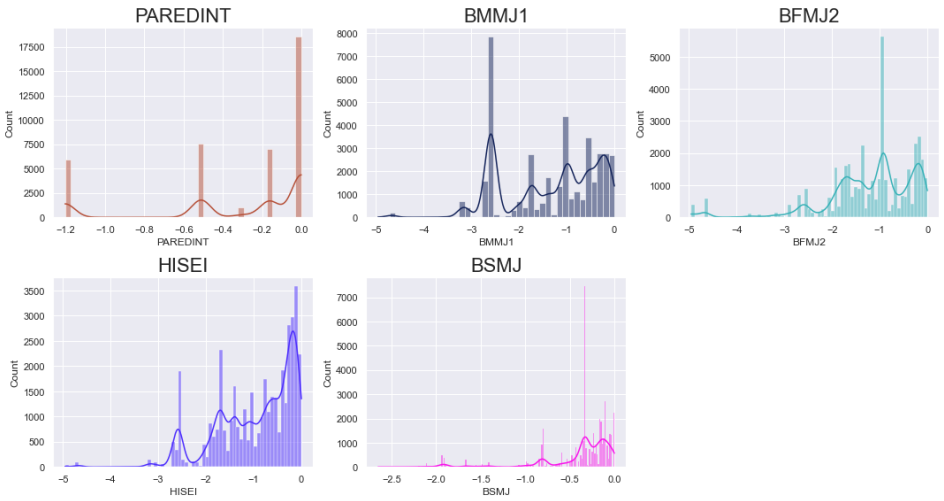


图 28: 取对数后 PAREDINT 等特征数据分布

5.3.3 组合特征：PARIDX

将上述特征相加得到 PARIDX 表示一个学生的父母地位和个人期望评判特征，因为每个特征的取值范围为 [0,1]，共 5 个特征所以新特征的取值范围为 [0,5]，组合后的效果比选用他们中的任意单个特征表现出的联系都要更强。

我们可以直观的看出，在复读的学生中，PARIDX 指数在 3 及以上的区间分布非常少，相比于不复读的学生，他们的 PARIDX 指数普遍都比较高，如图29：

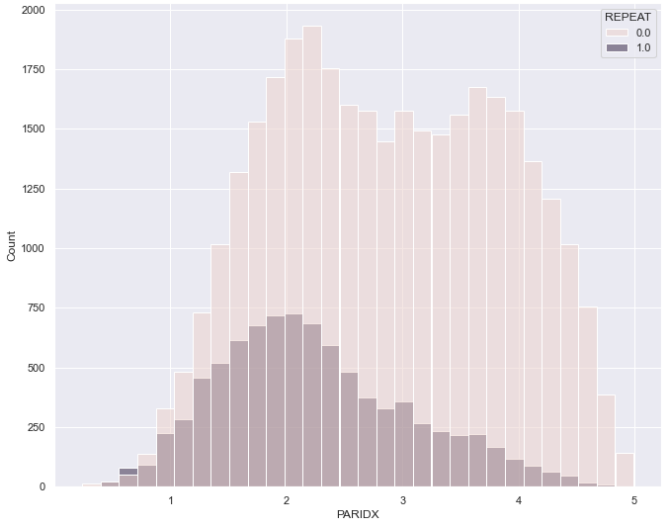


图 29: PARIDX 与 REPEAT 分布 (a)

通过图30可以更直观的看出，复读的学生与不复读的学生 PARIDX 密度分布的不同，复读的学生集中在低 PARIDX 指数区间。

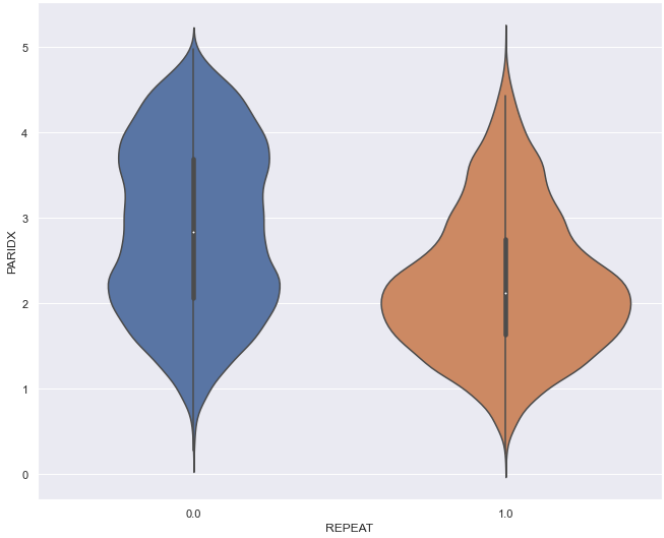
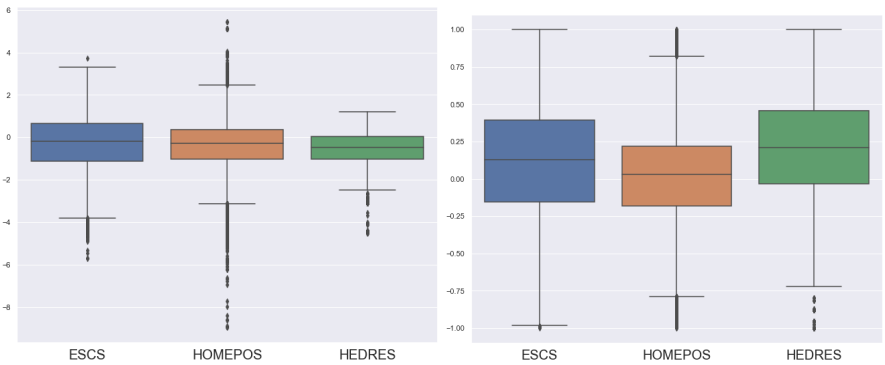


图 30: PARIDX 与 REPEAT 分布 (b)

5.4 FAMWEA: 家庭经济条件指标

5.4.1 异常值处理及归一化

观察数据的大致分布，其中箱线图的每个箱子都被压得很扁，说明受到较多的离群点影响，删除部分异常值后箱体变得更宽，如图31，在后续的柱形图中可以更直观的看出其变化。



(a) 初始 ESCS 等特征数据分布　　(b) 删除部分异常值后数据分布

图 31: ESCS 等特征数据分布 (a)

ESCS 和 HOMEPOS 的柱形图32显示它们接近正态分布，有一定偏度，先对异常值进行处理，因为异常值占比不高，所以选择将异常值删去。

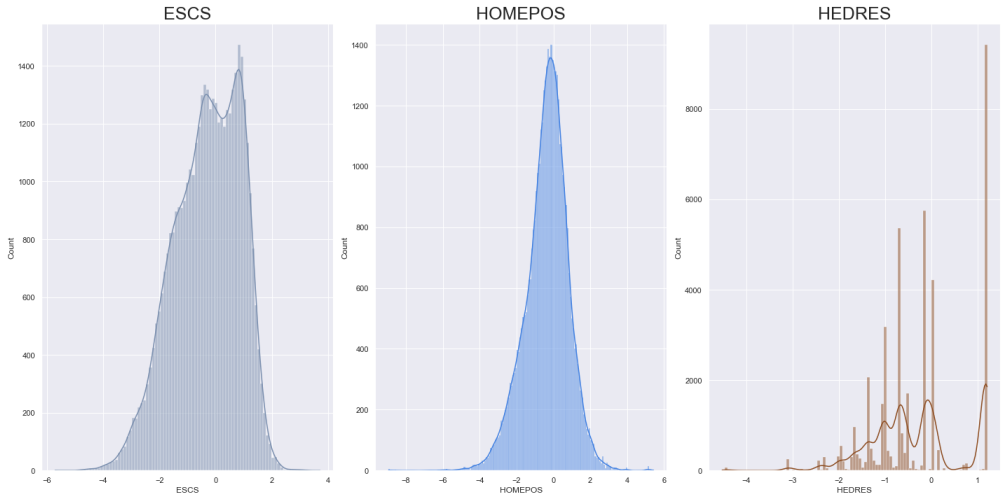


图 32: ESCS 等特征数据分布 (b)

再进行归一化，因为数据的取值可以为负数，所以将其归一化到 [-1,1] 区间上。删除异常值及归一化处理完后，离群点减少，得到的数据分布更均匀，图像“宽度”增加了，如图33。

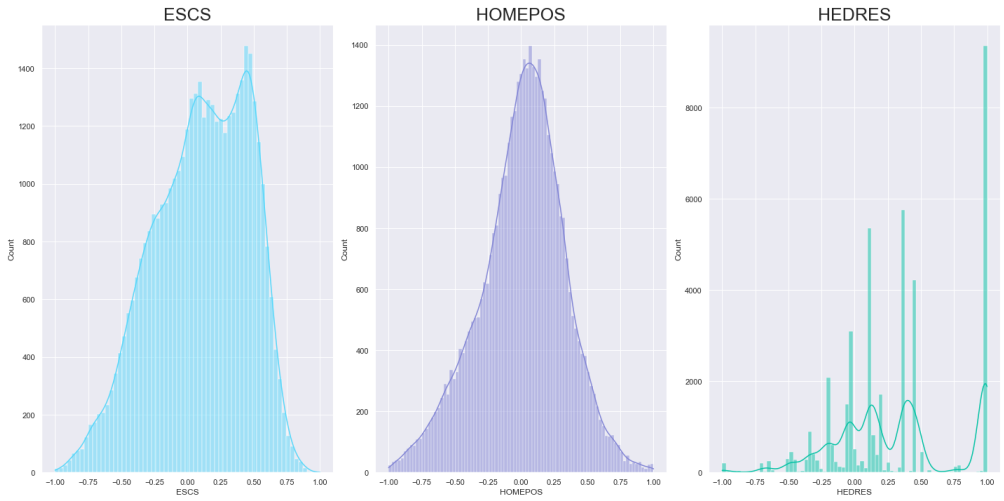


图 33: 处理后 ESCS 等特征数据分布

5.4.2 组合特征：FAMWEA

将上述 3 个特征相加 [5]，得到新的特征 FAMWEA 用以衡量一个学生家庭的经济条件，虽然相关系数并不高，但 FAMWEA 与 REPEAT 的密度分布呈现正态分布，并呈现一定正相关。

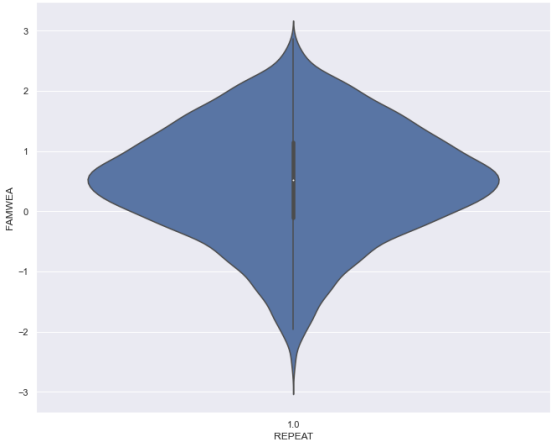


图 34: FAMWEA 与 REPEAT 密度分布

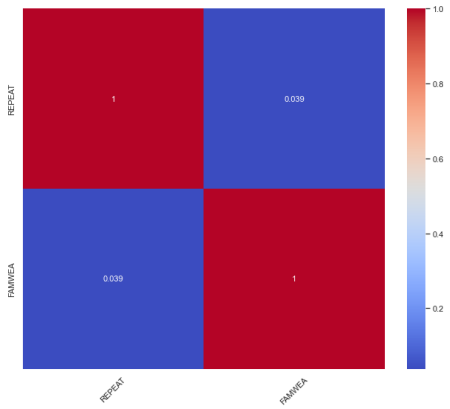


图 35: FAMWEA 与 REPEAT 相关性

6 总结

最终得到的特征如下，其中前 5 个特征是原始的单个特征，也是与 REPEAT 相关性最强的特征，后续特征是组合后得到的特征，组合特征中的单个特征也可以提出作为单个特征使用。其中 COUNT_1，PARIDX，Minor_SUM 表现出了与 REPEAT 较强的联系，其余特征也与 REPEAT 有一定的相关性。

表 7: Description of selected variables

特征	含义
ST001D01T	Student International Grade (Derived)
ST127Q01TA	Have you ever repeated a <grade>? At <ISCED 1>
ST127Q02TA	Have you ever repeated a <grade>? At <ISCED 2>
ST127Q03TA	Have you ever repeated a <grade>? At <ISCED 3>
GRADE	Grade compared to modal grade in country
PARIDX	父母地位及自我期望综合指标
COUNT_1	学生家庭学习配套硬件情况
FAMWEA	家庭经济条件指标
INTERTIM	学生使用互联网总时长
Math~Music_sumtime	学生分别在 Math、Music 等 9 门科目 (9 个特征) 上花费的总时间
Minor_SUM	学生在“主课”花费的学习时间
Major_SUM	学生在“副课”花费的学习时间

参考文献

[1] A. Mohd and S. Mohd, "A review of missing data treatment methods," Journal of Physics: Conference Series, vol. 1763, no. 1, p. 012011, 2021.

[2] D. Janzing, L. Minorics, and P. Bl"obaum, "Feature relevance quantification in explainable ai: A causal problem," in International Conference on Artificial Intelligence and Statistics. PMLR, 2020, pp. 2907–2916.

[3] World Bank, "Isced: International standard classification of education," [Accessed: 2023-05-01]. [Online]. Available: https://datatopics.worldbank.org/education/wbg/Classifications/ISCED-2011.html

[4] AvenueCyy, 数据挖掘：特征工程——特征处理与特征生成. 辽宁: CSDN, 2020.

[5] YYIverson, 【数据分析】特征工程中的特征构造、特征提取、特征选择. 上海: CSDN, 2023.

[6] V. Ottova, D. Hillebrandt, P. Brzoska, N. Dragano, S. Jordan, S. Muters, and H. Schroder, "Measurement and conceptualization of socio-economic status in the context of health inequality: a systematic review," International journal of public health, vol. 51, no. 5, pp. 347–359, 2006.

[7] Mystery, 数据预处理常用技巧 | 缺失值插补的 5 种方法. 上海: ZhiHU, 2023.