数据分析及实践 _Assignment3

Xiaoma

2023 年 5 月 15 日

1 实验要求

- 基于筛减版的 **PISA** 数据集，进行数据分析、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
- 特征抽取如
  - 特征的变换
  - 尝试组合特征
  - 特征子集选择

2 实验环境

VSCode + Python3.9.13

3 实验步骤

3.1 数据清洗

3.1.1 初步筛选特征

读取全部数据，并查看数据信息。该数据集共有 487 个特征，42176 条数据。
首先丢弃数据集中无意义的特征

- 索引特征：Unnamed:0 与 index
- 值唯一的特征：ADMINMODE 与 LANGTEST_COG

统计数据集中每个特征的缺失值数量，综合数据量以及特征量，将缺失值比率超过 0.1 的特征从数据集中丢弃，最终得到的特征数量为 179 个。

对照特征名与 **codebook** 表来推测每个特征对应的含义，通过比较发现，特征 CNTRYID 与 CNT，NatCen 意义重复，故丢弃特征 CNT,NatCen。特征 STRATUM，SUBNATIO 的地域信息更为详细，但数据过于复杂，故也被丢弃。

查看剩余特征的数据类型，均为 **int** 或 **float**，故不需要进行数据类型转换。

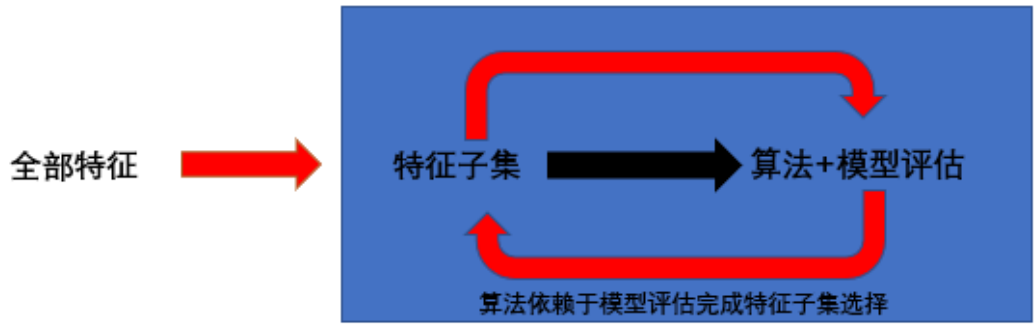
3.1.2 缺失值处理

计算数据集中每个特征的方差，发现特征的方差均较小，综合数据量以及特征量，使用均值填充缺失值。

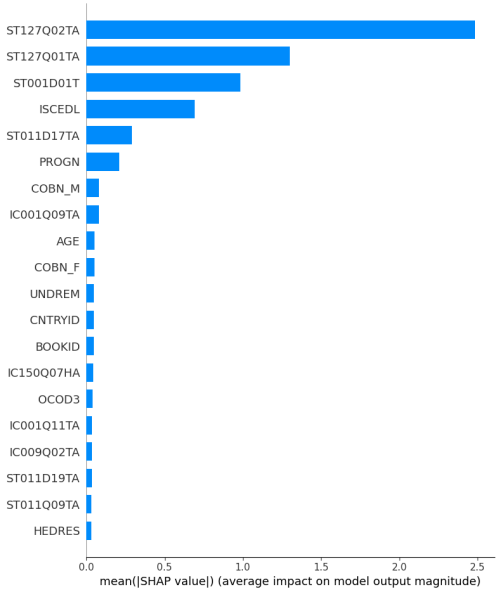
3.2 特征选择

3.2.1 基于嵌入法进行特征选择

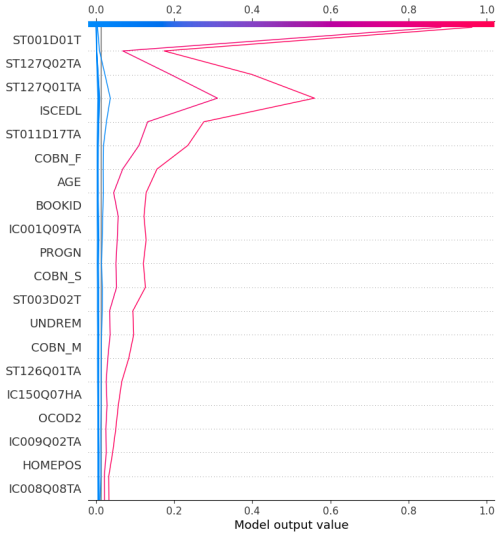
分别基于 **XGBClassifier,LGBMClassifier,DecisionTree** 对数据进行训练后，比较每个特征在预测过程中的重要性占比，可视化根据特征的预测过程，



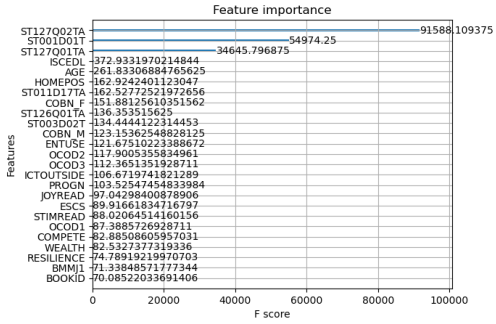
选择与预测任务最相关的若干特征绘制 pearson 相关系数热力图。特征重要性以及预测过程分别如图所示



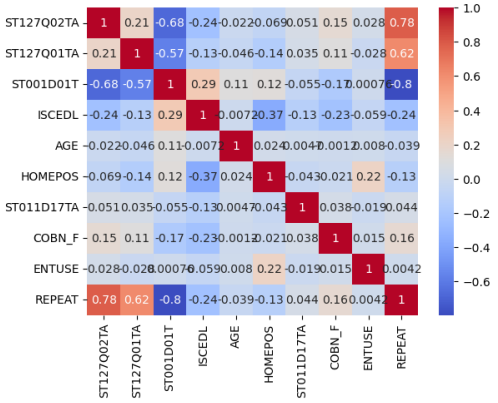
(a) *XGBoost_FeatureImportance*



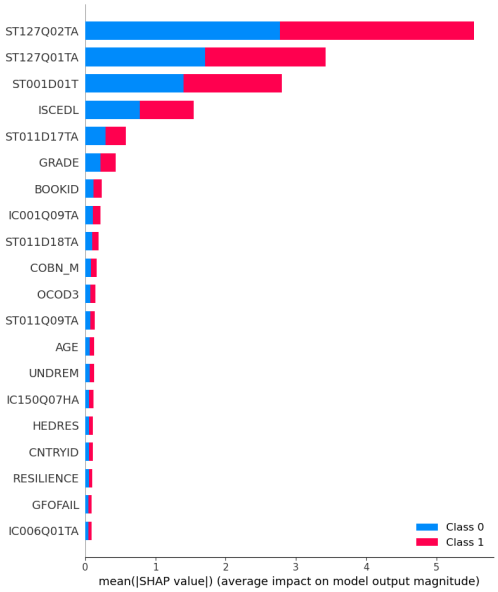
(b) *XGBoost_PredictProcess*



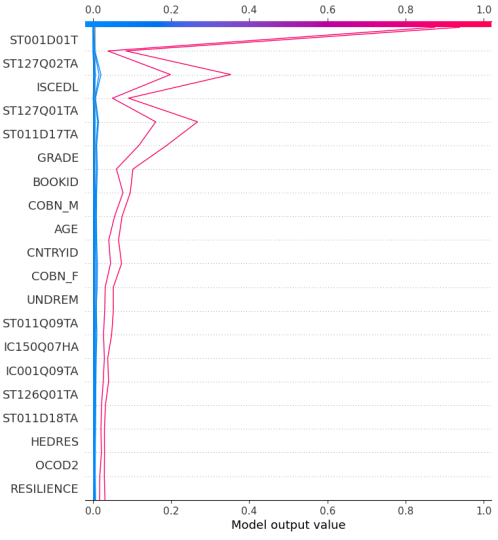
(c) *XGBoost_FeatureImportance*



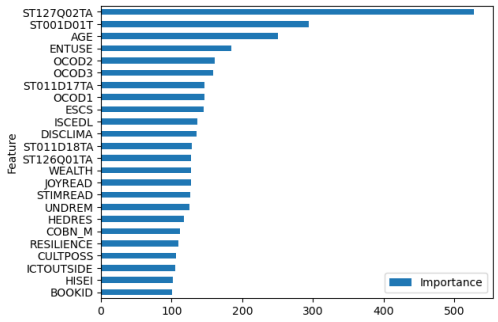
(d) *XGBoost_HeatMap*



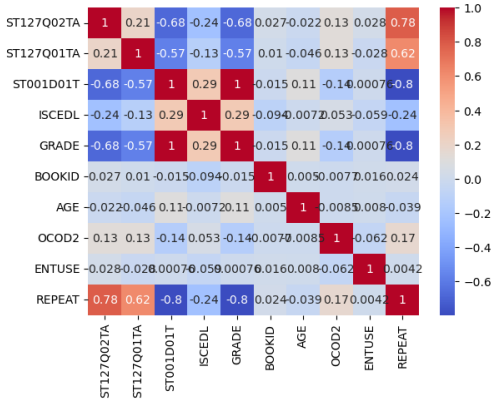
(e) *LGBM_FeatureImportance*



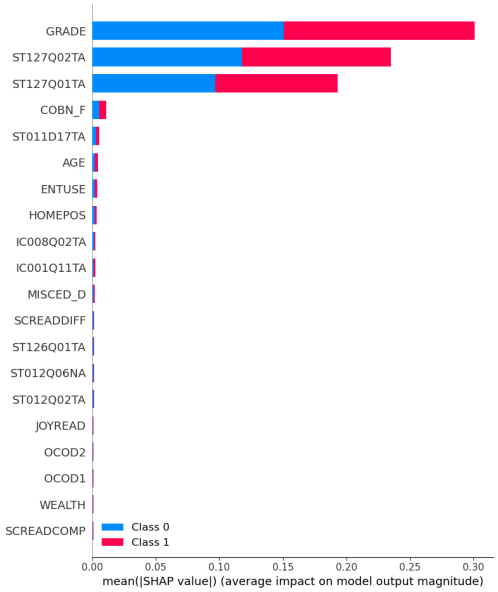
(f) *LGBM_PredictProcess*



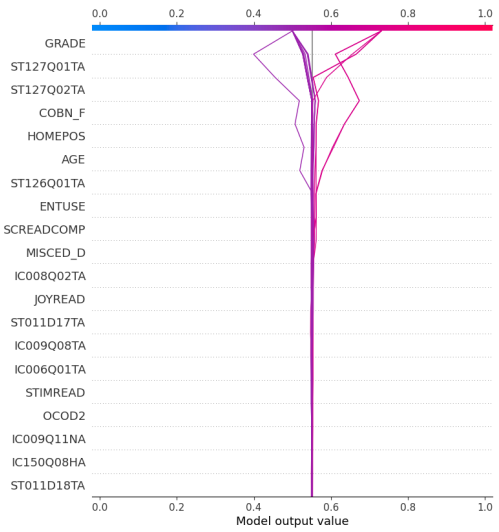
(g) *LGBM_FeatureImportance*



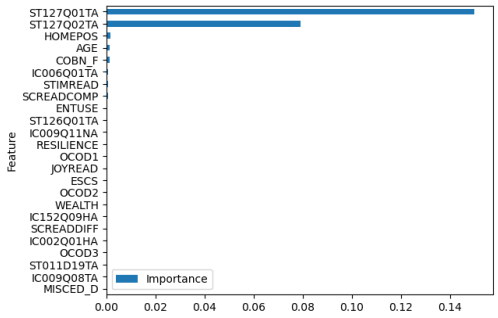
(h) *LGBM_HeatMap*



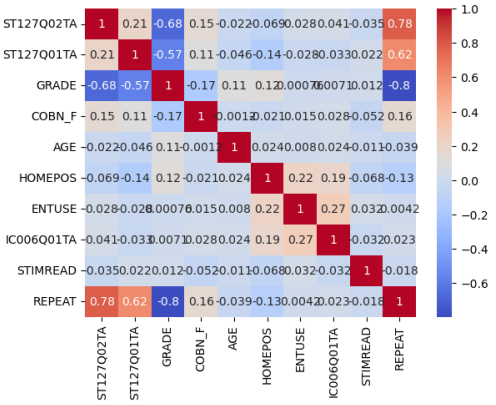
(i) *DecisionTree_FeatureImportance*



(j) *DecisionTree_PredictProcess*



(k) *DecisionTree_FeatureImportance*

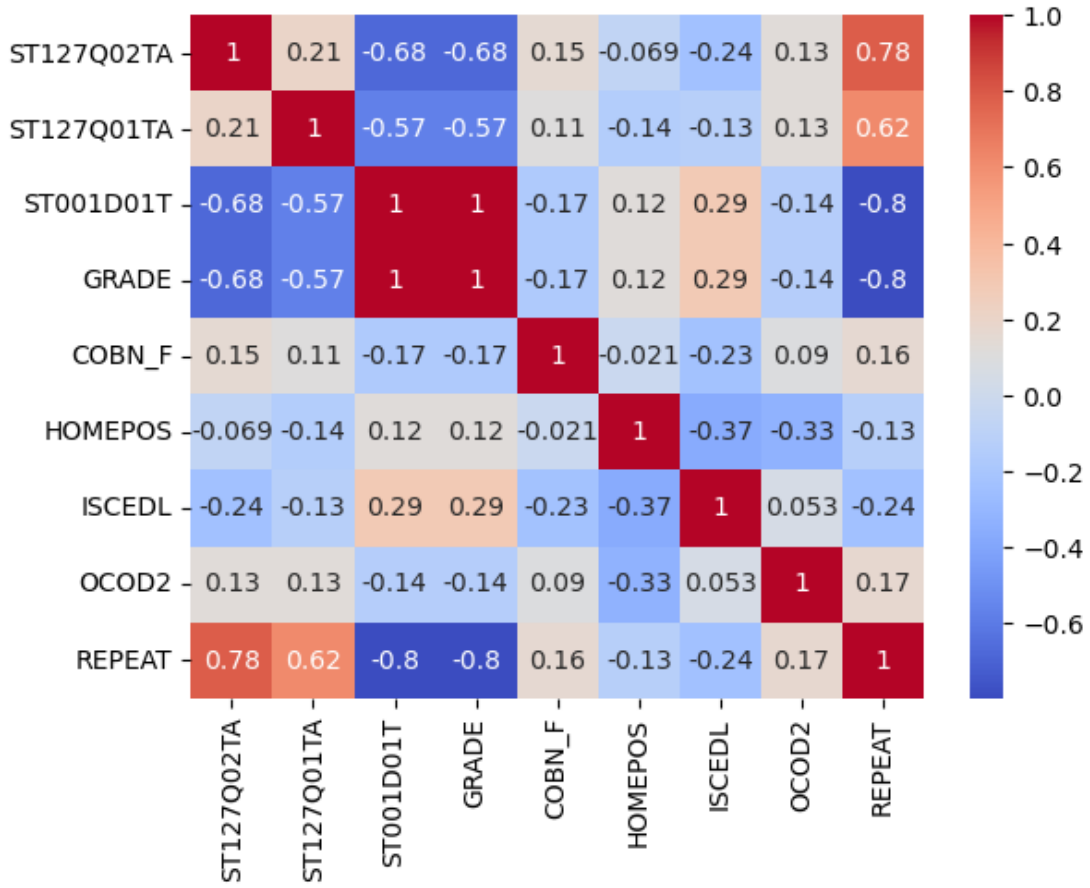


(l) *DecisionTree_HeatMap*

综合上述模型对应的特征重要性，最终选择 **ST127Q02TA,ST127Q01TA,ST001D01T,GRADE,COBN_F,HOMEPOS,ISCEDL,OCOD2** 作为 **REPEAT** 预测任务的训练特征。

3.3 相关性分析

所选择的 8 个特征与 **REPEAT** 的 pearson 相关系数为



分别使用 **XGBoost,DecisionTree** 对数据进行训练，10 折交叉验证的正确率分别为 **0.997,0.993**，与未进行数据筛选前精度几乎不变，意味着所选择的 8 个特征为参与预测任务的主要特征，丢弃其他特征对预测任务几乎无影响。

查询所选的 8 个特征对应的意义，并观察其值的分布规律。

通过查询 **codebook** 表，特征对应的意义分别为

特征名	含义
ST127Q02TA	Have you ever repeated a <grade>? At <ISCED 2>
ST127Q01TA	Have you ever repeated a <grade>? At <ISCED 1>
ST001D01T	Student International Grade (Derived)
GRADE	Grade compared to modal grade in country
COBN_F	Country of Birth National Categories- Father
HOMEPOS	Home possessions (WLE)
ISCEDL	ISCED level
OCOD2	ISCO-08 Occupation code - Father

4 总结

通过上述特征选择的过程，发现所选择的 8 个特征为预测 **REPEAT** 标签的重要特征，并且通过分析特征对应的含义发现，部分特征与预测任务的关系，根据生活经验仍可以判断具有强强相关性，并通过这些特征来发现剩余特征与预测任务的新的关系，说明本次特征选择是有意义且正确的。