数据处理分析实验报告IV

一． 数据集：PISA（国际学生评价计划）

一个衡量 15 岁学生能力水平的计划，分三个课程：数学，科学，阅读。

本次需要研究的是 18 年的评价结果，覆盖几十万名中学生的成绩。

二． 研究目的

1. 证明国家性别不平等比其他国家变量（国际基尼系数，国家人均 GDP）对学业成绩的影响更大；国家性别不平衡越严重，学生成绩越差
2. 证明成长型思维能缓解国家性别不平等的消极影响；研究成长型思维与国家性别不平衡的交互作用

三． 数据分析需求

这一次用 Lasso、随机森林、随机排序和 xgboost 来做特征重要性排名。

基于机器学习方法本身存在的不确定性，最终结果也是一种可能的情况，为分析结论做辅助补充。

四． 实验结果：

这一次实验只用到学校层和学生层，在控制变量中加入了 SNW

成绩用的是 pv2、pv5、pv9，数据总量约有 612004 条

经过清理之后剩余 444238 条数据

男生 221826 人，女生 222412 人

首先可以简单的看一下各个变量之间的相关性，容易看到 GII 和 GINI 都对成绩具有负面影响（颜色最深最黑）

4.1 Lasso 方法

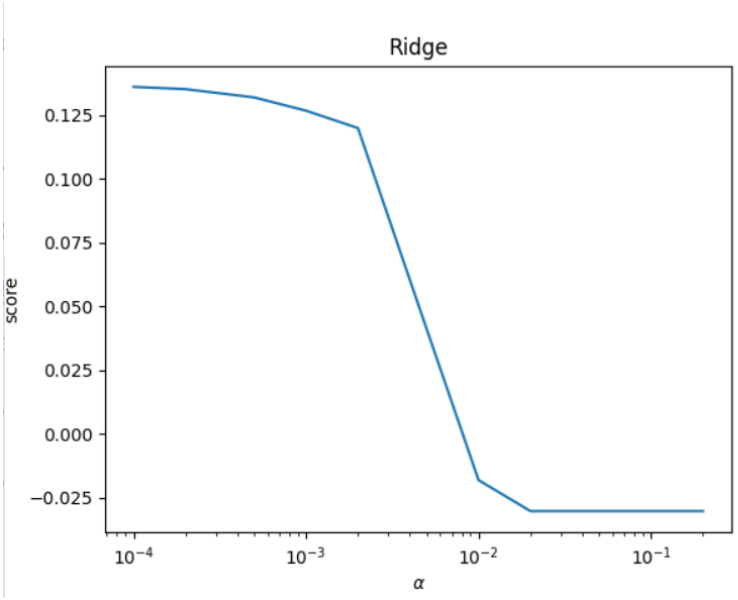
介绍

Lasoo 方法是一种正则化的方法，主要思想就是在经验风险上加一个惩罚项，这个惩罚项用于惩罚模型的复杂度，当惩罚项选用的是 L1 范数时就叫回归模型为 Lasso 回归。惩罚项的加入可以将一些不重要的变量前面的系数化为 0.

$$\widehat{\beta}_{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

因为在α越小，模型的评分（预测能力）越强，所以可以知道，七个特征变

量都对最终的成绩有一定的贡献。



当α从小取到大的过程中（α是 lasso 的参数），各个特征变量的系数开始收缩，（各个变量系数对应的名称从左到右依次为 gender、ESCS、gmc、SNW、gdp、GII、GINI）,因为模型会随着α的增大逐步舍弃次要的特征，系数变为 0 最慢的特征是最重要的特征，因此可以根据这个结果对特征变量的重要性进行排序:

#综合分析（男女混合在一起）（排序：重要的在前面）

实验 1.1.1-Lasso-PV2-math： GII、gender、gmc、ESCS、gdp 、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0002	Coe:-0.0305	0.2539	-0.0863	0.0082	-0.0793	0.0370
a:0.0010	Coe:-0.0273	0.2200	-0.0762	0.0000	-0.0831	0.0203
a:0.0050	Coe:-0.0110	0.0462	-0.0268	-0.0000	-0.0863	0.0000
a:0.0065	Coe:-0.0050	0.0000	-0.0073	-0.0000	-0.0799	0.0000
a:0.0077	Coe:-0.0002	0.0000	-0.0000	-0.0000	-0.0662	0.0000
a:0.0080	Coe:-0.0000	0.0000	-0.0000	-0.0000	-0.0627	0.0000

实验 1.1.2-Lasso-PV5-math：GII、gender、gmc、ESCS、gdp 、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0002	Coe:-0.0315	0.2633	-0.0894	0.0100	-0.0822	0.0382
a:0.0010	Coe:-0.0283	0.2293	-0.0794	0.0000	-0.0853	0.0218
a:0.0050	Coe:-0.0121	0.0556	-0.0299	-0.0000	-0.0892	0.0000
a:0.0065	Coe:-0.0060	0.0000	-0.0110	-0.0000	-0.0847	0.0000
a:0.0077	Coe:-0.0013	0.0000	-0.0000	-0.0000	-0.0713	0.0000
a:0.0080	Coe:-0.0000	0.0000	-0.0000	-0.0000	-0.0677	0.0000

实验 1.1.3-Lasso-PV9-math：GII、gender、gmc、ESCS、gdp 、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0002	Coe:-0.0315	0.2654	-0.0902	0.0105	-0.0832	0.0368
a:0.0010	Coe:-0.0283	0.2314	-0.0802	0.0000	-0.0861	0.0205
a:0.0050	Coe:-0.0121	0.0576	-0.0308	-0.0000	-0.0894	0.0000
a:0.0065	Coe:-0.0060	0.0000	-0.0120	-0.0000	-0.0854	0.0000
a:0.0077	Coe:-0.0013	0.0000	-0.0000	-0.0000	-0.0720	0.0000
a:0.0083	Coe:-0.0000	0.0000	-0.0000	-0.0000	-0.0649	0.0000

实验 1.1.4-Lasso-PV2-read：GII、ESCS、gmc、gender、Gini、gdp

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0002	Coe:0.0044	0.2385	-0.0657	-0.0420	-0.0931	0.0046
a:0.0005	Coe:0.0032	0.2265	-0.0611	-0.0179	-0.1027	0.0000
a:0.0010	Coe:0.0013	0.2052	-0.0543	-0.0000	-0.1079	0.0000
a:0.0050	Coe:0.0000	0.0309	-0.0044	-0.0000	-0.1019	0.0000
a:0.0055	Coe:0.0000	0.0087	-0.0000	-0.0000	-0.1011	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0701	0.0000

实验 1.1.5-Lasso-PV5-read：GII、ESCS、gmc、gender、Gini、gdp

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0002	Coe:0.0047	0.2549	-0.0708	-0.0476	-0.0978	0.0091
a:0.0005	Coe:0.0035	0.2433	-0.0660	-0.0218	-0.1100	0.0000
a:0.0010	Coe:0.0015	0.2221	-0.0591	-0.0000	-0.1165	0.0000
a:0.0035	Coe:0.0000	0.1132	-0.0279	-0.0000	-0.1128	0.0000
a:0.0060	Coe:0.0000	0.0035	-0.0000	-0.0000	-0.1090	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0827	0.0000

实验 1.1.6-Lasso-PV9-read：GII、ESCS、gmc、gender、Gini、gdp

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0002	Coe:0.0051	0.2443	-0.0659	-0.0427	-0.0932	0.0054
a:0.0005	Coe:0.0039	0.2323	-0.0613	-0.0183	-0.1033	0.0000
a:0.0010	Coe:0.0019	0.2110	-0.0545	-0.0000	-0.1085	0.0000
a:0.0035	Coe:0.0000	0.1022	-0.0232	-0.0000	-0.1048	0.0000
a:0.0055	Coe:0.0000	0.0146	-0.0000	-0.0000	-0.1018	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0721	0.0000

实验 1.1.7-Lasso-PV2-scie：GII、gmc、ESCS、Gdp、gender、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0001	Coe:-0.0016	0.2494	-0.0836	-0.0039	-0.0744	0.0300
a:0.0004	Coe:-0.0004	0.2370	-0.0796	-0.0000	-0.0786	0.0226
a:0.0010	Coe:-0.0000	0.2119	-0.0717	-0.0000	-0.0840	0.0091
a:0.0035	Coe:-0.0000	0.1037	-0.0402	-0.0000	-0.0845	0.0000
a:0.0065	Coe:-0.0000	0.0000	-0.0012	-0.0000	-0.0743	0.0000
a:0.0083	Coe:-0.0000	0.0000	-0.0000	-0.0000	-0.0533	0.0000

实验 1.1.8-Lasso-PV5- scie：GII、gmc、ESCS、Gdp、gender、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0001	Coe:-0.0020	0.2513	-0.0830	-0.0070	-0.0743	0.0304
a:0.0004	Coe:-0.0008	0.2389	-0.0789	-0.0000	-0.0798	0.0225
a:0.0010	Coe:-0.0000	0.2139	-0.0710	-0.0000	-0.0851	0.0090
a:0.0035	Coe:-0.0000	0.1056	-0.0396	-0.0000	-0.0856	0.0000
a:0.0065	Coe:-0.0000	0.0000	-0.0006	-0.0000	-0.0758	0.0000
a:0.0083	Coe:-0.0000	0.0000	-0.0000	-0.0000	-0.0547	0.0000

实验 1.1.9-Lasso-PV9- scie：GII、gmc、ESCS、Gdp、gender、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0001	Coe:-0.0022	0.2486	-0.0828	-0.0060	-0.0732	0.0310
a:0.0004	Coe:-0.0010	0.2362	-0.0787	-0.0000	-0.0783	0.0233
a:0.0010	Coe:-0.0000	0.2111	-0.0709	-0.0000	-0.0837	0.0097
a:0.0035	Coe:-0.0000	0.1029	-0.0394	-0.0000	-0.0844	0.0000
a:0.0065	Coe:-0.0000	0.0000	-0.0003	-0.0000	-0.0741	0.0000
a:0.0083	Coe:-0.0000	0.0000	-0.0000	-0.0000	-0.0530	0.0000

综上，可以看到性别不平等 GII 比 GINI 系数、GDP 变量的影响作用更大，并且因为系数是负值，所以其对成绩为负面影响。

#性别分析（男女生分开研究）：

男生成绩：

实验 1.2.1-Lasso-PV2-math：GII、gmc、ESCS、Gdp、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0001	Coe:0.0000	0.2478	-0.0876	0.0379	-0.0863	0.0383
a:0.0010	Coe:0.0000	0.2079	-0.0772	0.0000	-0.0786	0.0244
a:0.0035	Coe:0.0000	0.0975	-0.0462	-0.0000	-0.0851	0.0000
a:0.0065	Coe:0.0000	0.0000	-0.0079	-0.0000	-0.0717	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0508	0.0000

实验 1.2.2-Lasso-PV5-math：GII、gmc、ESCS、Gdp、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0001	Coe:0.0000	0.2573	-0.0907	0.0401	-0.0901	0.0396
a:0.0010	Coe:0.0000	0.2174	-0.0804	0.0000	-0.0815	0.0260
a:0.0035	Coe:0.0000	0.1071	-0.0494	-0.0000	-0.0888	0.0000
a:0.0065	Coe:0.0000	0.0000	-0.0116	-0.0000	-0.0773	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0566	0.0000

实验 1.2.3-Lasso-PV9-math：GII、gmc、ESCS、Gdp、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0001	Coe:0.0000	0.2627	-0.0930	0.0403	-0.0915	0.0388
a:0.0010	Coe:0.0000	0.2228	-0.0827	0.0000	-0.0828	0.0253
a:0.0035	Coe:0.0000	0.1125	-0.0517	-0.0000	-0.0897	0.0000
a:0.0065	Coe:0.0000	0.0000	-0.0142	-0.0000	-0.0793	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0587	0.0000

实验 1.2.4-Lasso-PV2-read：GII、gmc、ESCS、Gini、Gdp

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0001	Coe:0.0000	0.2374	-0.0693	-0.0249	-0.1027	0.0057
a:0.0003	Coe:0.0000	0.2291	-0.0663	-0.0077	-0.1106	0.0000
a:0.0035	Coe:0.0000	0.0853	-0.0273	-0.0000	-0.1069	0.0000
a:0.0055	Coe:0.0000	0.0000	-0.0029	-0.0000	-0.1021	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0690	0.0000

实验 1.2.5-Lasso-PV5-read：GII、gmc、ESCS、Gini、Gdp

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0001	Coe:0.0000	0.2580	-0.0773	-0.0349	-0.1088	0.0112
a:0.0005	Coe:0.0000	0.2415	-0.0713	-0.0005	-0.1245	0.0000
a:0.0035	Coe:0.0000	0.1065	-0.0350	-0.0000	-0.1188	0.0000
a:0.0060	Coe:0.0000	0.0000	-0.0043	-0.0000	-0.1128	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0857	0.0000

实验 1.2.6-Lasso-PV9-read：GII、gmc、ESCS、Gini、Gdp

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0001	Coe:0.0000	0.2426	-0.0706	-0.0275	-0.1024	0.0087
a:0.0004	Coe:0.0000	0.2302	-0.0661	-0.0017	-0.1143	0.0000
a:0.0035	Coe:0.0000	0.0908	-0.0285	-0.0000	-0.1089	0.0000
a:0.0057	Coe:0.0000	0.0000	-0.0014	-0.0000	-0.1030	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0721	0.0000

实验 1.2.7-Lasso-PV2-scie：GII、gmc、ESCS、Gdp、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0000	Coe:0.0000	0.2457	-0.0864	0.0008	-0.0759	0.0354
a:0.0004	Coe:0.0000	0.2289	-0.0814	-0.0000	-0.0789	0.0269
a:0.0035	Coe:0.0000	0.0917	-0.0431	-0.0000	-0.0853	0.0000
a:0.0057	Coe:0.0000	0.0000	-0.0161	-0.0000	-0.0796	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0496	0.0000

实验 1.2.8-Lasso-PV5- scie：GII、gmc、ESCS、Gdp、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0000	Coe:0.0000	0.2474	-0.0857	0.0013	-0.0771	0.0364
a:0.0004	Coe:0.0000	0.2305	-0.0808	-0.0000	-0.0798	0.0280
a:0.0035	Coe:0.0000	0.0935	-0.0425	-0.0000	-0.0868	0.0000
a:0.0057	Coe:0.0000	0.0000	-0.0155	-0.0000	-0.0814	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0514	0.0000

实验 1.2.9-Lasso-PV9- scie：GII、gmc、ESCS、Gdp、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0000	Coe:0.0000	0.2451	-0.0859	0.0021	-0.0757	0.0363
a:0.0004	Coe:0.0000	0.2282	-0.0810	-0.0000	-0.0781	0.0281
a:0.0035	Coe:0.0000	0.0911	-0.0427	-0.0000	-0.0852	0.0000
a:0.0057	Coe:0.0000	0.0000	-0.0156	-0.0000	-0.0793	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0493	0.0000

女生成绩：

实验 1.3.1-Lasso-PV2-math：GII、ESCS、gmc、Gdp、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0000	Coe:0.0000	0.2787	-0.0905	0.0214	-0.0892	0.0377
a:0.0004	Coe:0.0000	0.2629	-0.0859	0.0000	-0.0844	0.0320
a:0.0035	Coe:0.0000	0.1406	-0.0445	-0.0000	-0.0951	0.0000
a:0.0069	Coe:0.0000	0.0015	-0.0000	-0.0000	-0.0906	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0754	0.0000

实验 1.3.2-Lasso-PV5-math：GII、ESCS、gmc、Gdp、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0000	Coe:0.0000	0.2772	-0.0903	0.0222	-0.0880	0.0373
a:0.0004	Coe:0.0000	0.2613	-0.0857	0.0000	-0.0828	0.0318
a:0.0035	Coe:0.0000	0.1390	-0.0443	-0.0000	-0.0935	0.0000
a:0.0069	Coe:0.0000	0.0008	-0.0000	-0.0000	-0.0890	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0733	0.0000

实验 1.3.3-Lasso-PV9-math：GII、ESCS、gmc、Gdp、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0000	Coe:0.0000	0.2790	-0.0905	0.0233	-0.0896	0.0358
a:0.0004	Coe:0.0000	0.2630	-0.0859	0.0000	-0.0839	0.0304
a:0.0035	Coe:0.0000	0.1406	-0.0446	-0.0000	-0.0939	0.0000
a:0.0069	Coe:0.0000	0.0016	-0.0000	-0.0000	-0.0894	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0742	0.0000

实验 1.3.4-Lasso-PV2-read：GII、ESCS、gmc、Gini、Gdp

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0000	Coe:0.0000	0.2525	-0.0683	-0.0868	-0.0721	0.0148
a:0.0004	Coe:0.0000	0.2382	-0.0614	-0.0510	-0.0901	0.0000
a:0.0035	Coe:0.0000	0.1156	-0.0201	-0.0000	-0.1028	0.0000
a:0.0060	Coe:0.0000	0.0117	-0.0000	-0.0000	-0.0993	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0754	0.0000

实验 1.3.5-Lasso-PV5-read：GII、ESCS、gmc、Gini、Gdp

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0000	Coe:0.0000	0.2627	-0.0702	-0.0866	-0.0756	0.0181
a:0.0005	Coe:0.0000	0.2456	-0.0619	-0.0437	-0.0974	0.0000
a:0.0035	Coe:0.0000	0.1260	-0.0219	-0.0000	-0.1077	0.0000
a:0.0060	Coe:0.0000	0.0226	-0.0000	-0.0000	-0.1043	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0829	0.0000

实验 1.3.6-Lasso-PV9-read：GII、ESCS、gmc、Gini、Gdp

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0000	Coe:0.0000	0.2686	-0.0698	-0.0888	-0.0753	0.0138
a:0.0005	Coe:0.0000	0.2512	-0.0617	-0.0474	-0.0946	0.0000
a:0.0035	Coe:0.0000	0.1317	-0.0216	-0.0000	-0.1062	0.0000
a:0.0060	Coe:0.0000	0.0282	-0.0000	-0.0000	-0.1028	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0826	0.0000

实验 1.3.7-Lasso-PV2-scie：GII、ESCS、gmc、Gdp、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0000	Coe:0.0000	0.2691	-0.0880	-0.0285	-0.0670	0.0337
a:0.0005	Coe:0.0000	0.2515	-0.0802	-0.0000	-0.0830	0.0178
a:0.0035	Coe:0.0000	0.1320	-0.0407	-0.0000	-0.0873	0.0000
a:0.0067	Coe:0.0000	0.0029	-0.0000	-0.0000	-0.0832	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0653	0.0000

实验 1.3.7-Lasso-PV5-scie：GII、ESCS、gmc、Gdp、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0000	Coe:0.0000	0.2668	-0.0860	-0.0348	-0.0646	0.0328
a:0.0005	Coe:0.0000	0.2495	-0.0780	-0.0000	-0.0832	0.0158
a:0.0035	Coe:0.0000	0.1299	-0.0386	-0.0000	-0.0866	0.0000
a:0.0067	Coe:0.0000	0.0002	-0.0000	-0.0000	-0.0824	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0639	0.0000

实验 1.3.7-Lasso-PV9-scie：GII、ESCS、gmc、Gdp、Gini

alpha	gender	ESCS	gmc	Gini	GII	Gdp
a:0.0000	Coe:0.0000	0.2688	-0.0869	-0.0343	-0.0649	0.0348
a:0.0005	Coe:0.0000	0.2515	-0.0789	-0.0000	-0.0833	0.0179
a:0.0035	Coe:0.0000	0.1320	-0.0394	-0.0000	-0.0877	0.0000
a:0.0067	Coe:0.0000	0.0026	-0.0000	-0.0000	-0.0835	0.0000
a:0.0083	Coe:0.0000	0.0000	-0.0000	-0.0000	-0.0656	0.0000

综上，可以观察到 GII 的重要性排名在第一位置，其次是 ESCS。

可证明，GII 对成绩的影响程度比其他国家变量更深。

4.2 随机森林

介绍

随机森林（Random Forest，简称RF）属于集成学习方法，是Bagging方法的一个扩展变体. RF在以决策树为基学习器构建Bagging集成的基础上，进一步在决策树的训练过程中引入了随机属性选择，具体来说，传统决策树在选择划分属性时是在当前结点的属性集合（假定有d个属性）中选择一个最优属性；而在RF中，对基决策树的每个结点，先从该结点的属性集合中随机选择一个包含k个属性的子集，然后再从这个子集中选择一个最优属性用于划分.

随机森林通常用于数据科学工作流程中的特征选择。 原因是因为随机森林使用的基于树的策略自然会根据它们提高节点purity的程度来排名。 这意味着所有树的impurity减少（称为gini impurity）。 impurity减少最多的节点出现在树的开头，而impurity减少最少的节点出现在树的末尾。 因此，通过修剪特定节点下方的树，我们可以得到最重要特征的子集。

#综合实验：

```
实验-randomforest-因变量为 math2 时
[('ESCS', 0.7204), ('gmc', 0.0772), ('Gdp', 0.0698), ('GII', 0.0674), ('gini', 0.046), ('gender', 0.0191)]
实验-randomforest-因变量为 math5 时
[('ESCS', 0.7226), ('gmc', 0.077), ('GII', 0.0741), ('Gdp', 0.0606), ('gini', 0.0465), ('gender', 0.0193)]
实验-randomforest-因变量为 math9 时
[('ESCS', 0.7215), ('gmc', 0.0772), ('Gdp', 0.0715), ('GII', 0.0647), ('gini', 0.0458), ('gender', 0.0192)]
实验-randomforest-因变量为 read2 时
[('ESCS', 0.7338), ('GII', 0.1029), ('gmc', 0.0596), ('gini', 0.0419), ('Gdp', 0.0354), ('gender', 0.0263)]
实验-randomforest-因变量为 read5 时
[('ESCS', 0.7341), ('GII', 0.104), ('gmc', 0.0587), ('gini', 0.0426), ('Gdp', 0.034), ('gender', 0.0266)]
实验-randomforest-因变量为 read9 时
[('ESCS', 0.7369), ('GII', 0.1007), ('gmc', 0.0589), ('gini', 0.0429), ('Gdp', 0.0349), ('gender', 0.0259)]
实验-randomforest-因变量为 scie2 时
[('ESCS', 0.7285), ('Gdp', 0.0762), ('gmc', 0.0686), ('GII', 0.0602), ('gini', 0.0337), ('gender', 0.0328)]
实验-randomforest-因变量为 scie5 时
[('ESCS', 0.73), ('Gdp', 0.0793), ('gmc', 0.0675), ('GII', 0.0552), ('gini', 0.0348), ('gender', 0.0331)]
实验-randomforest-因变量为 scie9 时
[('ESCS', 0.7286), ('Gdp', 0.071), ('gmc', 0.0693), ('GII', 0.0621), ('gini', 0.0367), ('gender', 0.0322)]
```

上图结果展示的是数学成绩 pv259 和阅读成绩 pv259 作为因变量时的变量重要性排名，可以看到 GII 排名在第二三四。（对于变量的特征相对重要性的得分是利用模型提供的 feature_importance_属性得到的）

#性别分析（男女分别研究）

男生成绩：

```
实验-randomforest-因变量为 math2 时
[('ESCS', 0.7361), ('gmc', 0.0748), ('Gdp', 0.0704), ('GII', 0.0697), ('gini', 0.049), ('gender', 0.0)]
实验-randomforest-因变量为 math5 时
[('ESCS', 0.7359), ('GII', 0.076), ('gmc', 0.0751), ('Gdp', 0.0639), ('gini', 0.0492), ('gender', 0.0)]
实验-randomforest-因变量为 math9 时
[('ESCS', 0.7364), ('gmc', 0.0749), ('GII', 0.0711), ('Gdp', 0.0685), ('gini', 0.0492), ('gender', 0.0)]
实验-randomforest-因变量为 read2 时
[('ESCS', 0.7431), ('GII', 0.1181), ('gmc', 0.0601), ('gini', 0.0412), ('Gdp', 0.0376), ('gender', 0.0)]
实验-randomforest-因变量为 read5 时
[('ESCS', 0.7412), ('GII', 0.1197), ('gmc', 0.0605), ('gini', 0.0414), ('Gdp', 0.0372), ('gender', 0.0)]
实验-randomforest-因变量为 read9 时
[('ESCS', 0.7432), ('GII', 0.1188), ('gmc', 0.0603), ('gini', 0.0405), ('Gdp', 0.0372), ('gender', 0.0)]
实验-randomforest-因变量为 scie2 时
[('ESCS', 0.7395), ('Gdp', 0.0783), ('gmc', 0.0712), ('GII', 0.0705), ('gini', 0.0405), ('gender', 0.0)]
实验-randomforest-因变量为 scie5 时
[('ESCS', 0.7398), ('Gdp', 0.0791), ('gmc', 0.0703), ('GII', 0.0696), ('gini', 0.0412), ('gender', 0.0)]
实验-randomforest-因变量为 scie9 时
[('ESCS', 0.7365), ('Gdp', 0.0808), ('gmc', 0.0719), ('GII', 0.0692), ('gini', 0.0416), ('gender', 0.0)]
```

女生成绩：

```
实验-randomforest-因变量为 math2 时
[('ESCS', 0.7444), ('GII', 0.0835), ('gmc', 0.0757), ('Gdp', 0.0595), ('gini', 0.0369), ('gender', 0.0)]
实验-randomforest-因变量为 math5 时
[('ESCS', 0.7479), ('GII', 0.0833), ('gmc', 0.0752), ('Gdp', 0.0565), ('gini', 0.0372), ('gender', 0.0)]
实验-randomforest-因变量为 math9 时
[('ESCS', 0.7484), ('GII', 0.0823), ('gmc', 0.0759), ('Gdp', 0.0566), ('gini', 0.0368), ('gender', 0.0)]
实验-randomforest-因变量为 read2 时
[('ESCS', 0.7455), ('GII', 0.1029), ('gmc', 0.0629), ('gini', 0.0526), ('Gdp', 0.0362), ('gender', 0.0)]
实验-randomforest-因变量为 read5 时
[('ESCS', 0.7478), ('GII', 0.1026), ('gmc', 0.0613), ('gini', 0.0521), ('Gdp', 0.0361), ('gender', 0.0)]
实验-randomforest-因变量为 read9 时
[('ESCS', 0.7492), ('GII', 0.1004), ('gmc', 0.0621), ('gini', 0.0523), ('Gdp', 0.036), ('gender', 0.0)]
实验-randomforest-因变量为 scie2 时
[('ESCS', 0.7522), ('GII', 0.0822), ('gmc', 0.0718), ('Gdp', 0.0497), ('gini', 0.0441), ('gender', 0.0)]
实验-randomforest-因变量为 scie5 时
[('ESCS', 0.754), ('GII', 0.0809), ('gmc', 0.0706), ('Gdp', 0.0492), ('gini', 0.0453), ('gender', 0.0)]
实验-randomforest-因变量为 scie9 时
[('ESCS', 0.7532), ('GII', 0.0822), ('gmc', 0.0717), ('Gdp', 0.0482), ('gini', 0.0448), ('gender', 0.0)]
```

上图结果展示的是分别计算男女成绩，数学成绩 pv259 和阅读成绩 pv259 作为因变量时的变量重要性排名，可以看到 GII 排名在第二时比较多。且 GII 的影响对于女生比对于男生影响程度更多。

4.3 Xgboost 方法

介绍

XGBoost 是陈天奇等人开发的一个开源机器学习项目，高效地实现了 GBDT 算法并进行了算法和工程上的许多改进，被广泛应用在 Kaggle 竞赛及其他许多机器学习竞赛中并取得了不错的成绩。说到 XGBoost，不得不提 GBDT(Gradient Boosting Decision Tree)。因为 XGBoost 本质上还是一个 GBDT，但是力争把速度和效率发挥到极致，所以叫 X (Extreme) GBoosted。包括前面说过，两者都是 boosting 方法。

xgboost 属于决策树类的算法，决策树具备评估特征重要性的能力。单个决策树的重要性是通过每个属性分割点（树的分叉点）改进性能度量的量计算的，由节点负责的观察数加权。因此 xgboost 也能用于特征重要性评估。

#综合实验:

```
实验-xgboost-因变量为 math2 时
[('gmc', 0.23981017), ('Gdp', 0.20023079), ('GII', 0.1792558), ('ESCS', 0.15434647), ('gender', 0.1376622), ('Gini', 0.08869452)]
实验-xgboost-因变量为 math5 时
[('gmc', 0.22988242), ('GII', 0.18754523), ('Gdp', 0.1754084), ('gender', 0.15818864), ('ESCS', 0.15784082), ('Gini', 0.09113452)]
实验-xgboost-因变量为 math9 时
[('gmc', 0.23299293), ('Gdp', 0.19665053), ('gender', 0.16788438), ('ESCS', 0.15630154), ('GII', 0.15322821), ('Gini', 0.09294234)]
实验-xgboost-因变量为 read2 时
[('GII', 0.371706), ('ESCS', 0.20086859), ('gmc', 0.18150647), ('Gdp', 0.123435974), ('Gini', 0.10126352), ('gender', 0.021219425)]
实验-xgboost-因变量为 read5 时
[('GII', 0.3322895), ('gmc', 0.19113825), ('ESCS', 0.19038743), ('Gini', 0.13765666), ('Gdp', 0.12921937), ('gender', 0.019308839)]
实验-xgboost-因变量为 read9 时
[('GII', 0.3274635), ('ESCS', 0.19656043), ('gmc', 0.1783091), ('Gdp', 0.13970025), ('Gini', 0.13454378), ('gender', 0.023422915)]
实验-xgboost-因变量为 scie2 时
[('Gdp', 0.25507528), ('gmc', 0.24899858), ('GII', 0.216881), ('ESCS', 0.1822027), ('Gini', 0.07792921), ('gender', 0.018913211)]
实验-xgboost-因变量为 scie5 时
[('gmc', 0.2509025), ('Gdp', 0.23830523), ('GII', 0.20107517), ('ESCS', 0.19864357), ('Gini', 0.09475482), ('gender', 0.016318718)]
实验-xgboost-因变量为 scie9 时
[('gmc', 0.24961348), ('Gdp', 0.23590498), ('GII', 0.20728466), ('ESCS', 0.18480746), ('Gini', 0.101556435), ('gender', 0.020832982)]
```

#性别分析（男女分别研究）

```
男生成绩
实验-xgboost-因变量为 math2 时
[('gmc', 0.27222118), ('GII', 0.2371495), ('Gdp', 0.22113031), ('ESCS', 0.1555129), ('Gini', 0.11398612), ('gender', 0.0)]
实验-xgboost-因变量为 math5 时
[('gmc', 0.2716958), ('GII', 0.24961755), ('Gdp', 0.20425695), ('ESCS', 0.15998109), ('Gini', 0.11444864), ('gender', 0.0)]
实验-xgboost-因变量为 math9 时
[('gmc', 0.27551883), ('GII', 0.22852343), ('Gdp', 0.21886027), ('ESCS', 0.1532673), ('Gini', 0.12383021), ('gender', 0.0)]
实验-xgboost-因变量为 read2 时
[('GII', 0.4218317), ('gmc', 0.18105681), ('ESCS', 0.16052176), ('Gdp', 0.13903023), ('Gini', 0.09755943), ('gender', 0.0)]
实验-xgboost-因变量为 read5 时
[('GII', 0.40903008), ('gmc', 0.18302819), ('ESCS', 0.17152418), ('Gdp', 0.13364008), ('Gini', 0.10277743), ('gender', 0.0)]
实验-xgboost-因变量为 read9 时
[('GII', 0.43164724), ('gmc', 0.17446171), ('ESCS', 0.16296916), ('Gdp', 0.12781411), ('Gini', 0.10310773), ('gender', 0.0)]
实验-xgboost-因变量为 scie2 时
[('GII', 0.2601484), ('Gdp', 0.25765482), ('gmc', 0.23221599), ('ESCS', 0.1552338), ('Gini', 0.09474706), ('gender', 0.0)]
实验-xgboost-因变量为 scie5 时
[('Gdp', 0.2546339), ('gmc', 0.25154686), ('GII', 0.24034327), ('ESCS', 0.15320018), ('Gini', 0.100275785), ('gender', 0.0)]
实验-xgboost-因变量为 scie9 时
[('Gdp', 0.2577719), ('GII', 0.25361744), ('gmc', 0.24621095), ('ESCS', 0.15186705), ('Gini', 0.09053264), ('gender', 0.0)]
```

```
女生成绩
实验-xgboost-因变量为 math2 时
[('GII', 0.27902186), ('gmc', 0.24797592), ('Gdp', 0.20626536), ('ESCS', 0.19089098), ('Gini', 0.07584583), ('gender', 0.0)]
实验-xgboost-因变量为 math5 时
[('GII', 0.2791944), ('gmc', 0.2674884), ('ESCS', 0.20489758), ('Gdp', 0.17975876), ('Gini', 0.06866097), ('gender', 0.0)]
实验-xgboost-因变量为 math9 时
[('gmc', 0.2904919), ('GII', 0.28160003), ('ESCS', 0.1998096), ('Gdp', 0.1536584), ('Gini', 0.07444007), ('gender', 0.0)]
实验-xgboost-因变量为 read2 时
[('GII', 0.36776808), ('ESCS', 0.19499001), ('gmc', 0.18636368), ('Gdp', 0.13404049), ('Gini', 0.11683767), ('gender', 0.0)]
实验-xgboost-因变量为 read5 时
[('GII', 0.37996387), ('ESCS', 0.19550884), ('gmc', 0.17163624), ('Gini', 0.1309674), ('Gdp', 0.12192361), ('gender', 0.0)]
实验-xgboost-因变量为 read9 时
[('GII', 0.34218407), ('ESCS', 0.20754306), ('gmc', 0.18598886), ('Gini', 0.13796479), ('Gdp', 0.12631926), ('gender', 0.0)]
实验-xgboost-因变量为 scie2 时
[('GII', 0.29365653), ('gmc', 0.25172505), ('ESCS', 0.19860359), ('Gdp', 0.15839027), ('Gini', 0.09762452), ('gender', 0.0)]
实验-xgboost-因变量为 scie5 时
[('GII', 0.2885254), ('gmc', 0.24590701), ('ESCS', 0.19953543), ('Gdp', 0.15887499), ('Gini', 0.10715719), ('gender', 0.0)]
实验-xgboost-因变量为 scie9 时
[('GII', 0.28499794), ('gmc', 0.24582596), ('ESCS', 0.19342023), ('Gdp', 0.17054056), ('Gini', 0.10521531), ('gender', 0.0)]
```

五． 实验结论：

本实验通过 Lasso 参数选择方法、随机森林方法、xgboost 方法，从多个角度，以不同的评估方法衡量 ESCS、gender、gmc、GII、gdp 、GINI 这些自变量对学生成绩的影响，虽然从不同角度评估各个自变量重要性时重要性排名有所差别，但是国家性别不平等在其中都有着较大的影响。证明了国家性别不平等比其他国家变量（国际基尼系数，国家人均 GDP）对学业成绩的影响更大；国家性别不平衡越严重，学生成绩越差。此外，我们同样

证明了国家性别不平等对男女学生成绩都有一定的负面影响。