

美国高校招生填报率统计数据的分析和预测

中山大学 MA7259《机器学习》期中作业

何瑞杰

2025-11-14

中山大学 | 大湾区大学

目录

1. 研究背景与目的	1
2. 探索性数据分析	3
2.1 重要统计量分析	5
2.2 相关矩阵	7
2.3 公立学校和私立学校的部分特征数据分析	9
2.4 降维可视化	13
2.5 正态性检验	14
3. 方法和实验	15
3.1 数据预处理	16
3.2 线性回归及其变体	17
3.3 决策树	24
4. 结论和讨论	27

1. 研究背景与目的

高等学校的填报是国内高中毕业生在结束高考后面临的第一重难题，在大洋彼岸也是如此。在择校时，我们常常会事先了解学校所处的地域、办学能力、师资、生源、学生毕业去向、校园生活开销等等信息后再做决定。我们想知道这些因素是如何影响填报的选择，或者说是学校收到的申请数量？

和中国大陆不同的是，美国的学生可以同时投递多所高校，但其投递的偏好依旧反映了学校的特征。本课题研究采用开放的 1995 年美国高校数据集 college 尝试通过是否为私立学校、实际入学人数、本科生人数、食宿费用等等对其报名人数进行预测，并观察在预测过程中起重要作用的因素。

目录

1. 研究背景与目的	1
2. 探索性数据分析	3
2.1 重要统计量分析	5
2.2 相关矩阵	7
2.3 公立学校和私立学校的部分特征数据分析	9
2.4 降维可视化	13
2.5 正态性检验	14
3. 方法和实验	15
3.1 数据预处理	16
3.2 线性回归及其变体	17
3.3 决策树	24
4. 结论和讨论	27

2. 探索性数据分析

表 1 college 数据集的各数据域名称及含义

数据域名称	含义	数据类型
P.Undergrad	非全日制本科学生人数	int
Outstate	外州学生学费	int
Room.Board	食宿费用	int
Books	预估书本费用	int
Personal	预估个人开销	int
PhD	拥有博士学位的教师比例	int
Terminal	拥有最高学位的教师比例	int
S.F.Ratio	师生比例	float
perc.alumni	捐赠校友比例	int
Expend	生均教学支出	int
Grad.Rate	毕业率	int

2.1 重要统计量分析

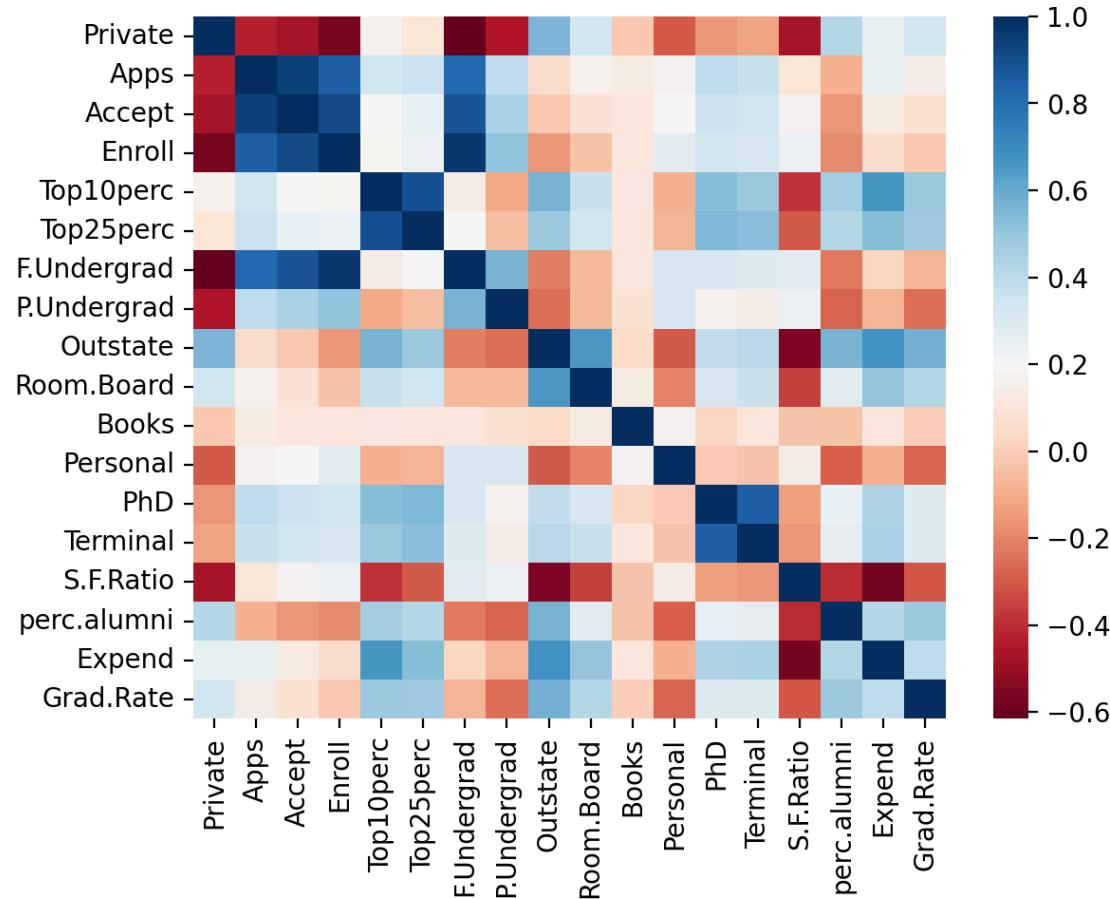
表 2 college 数据集的各数据域名称及含义

数据域	均值	标准差	最小值	25% 分位数	中位数	75% 分位数	最大值
Private	0.72	0.44	0.0	0.0	1.0	1.0	1.0
Apps	3001.63	3870.20	81.0	776.0	1558.0	3624.0	48094.0
Top10perc	27.55	17.64	1.0	15.0	23.0	35.0	96.0
Top25perc	55.79	19.80	9.0	41.0	54.0	69.0	100.0
F.Undergrad	3699.90	4850.42	139.0	992.0	1707.0	4005.0	31643.0
Outstate	10440.66	4023.01	2340.0	7320.0	9990.0	12925.0	21700.0
Books	549.38	165.10	96.0	470.0	500.0	600.0	2340.0
Personal	1340.64	677.07	250.0	850.0	1200.0	1700.0	6800.0
PhD	72.66	16.32	8.0	62.0	75.0	85.0	103.0
S.F.Ratio	14.08	3.95	2.5	11.5	13.6	16.5	39.8
perc.alumni	22.74	12.39	0.0	13.0	21.0	31.0	64.0
Expend	9660.17	5221.76	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	65.44	17.11	10.0	53.0	65.0	78.0	100.0

2.1 重要统计量分析

1. 数据集以私立学校为主 (72.7%)
2. 大学在规模和资源上存在明显的头部效应
3. 从申请到录取再到入学，数量大幅减少，平均入学率仅为 38.6%
5. 不同学校的师生比、师资博士比例、基础生活成本等指标差异不大

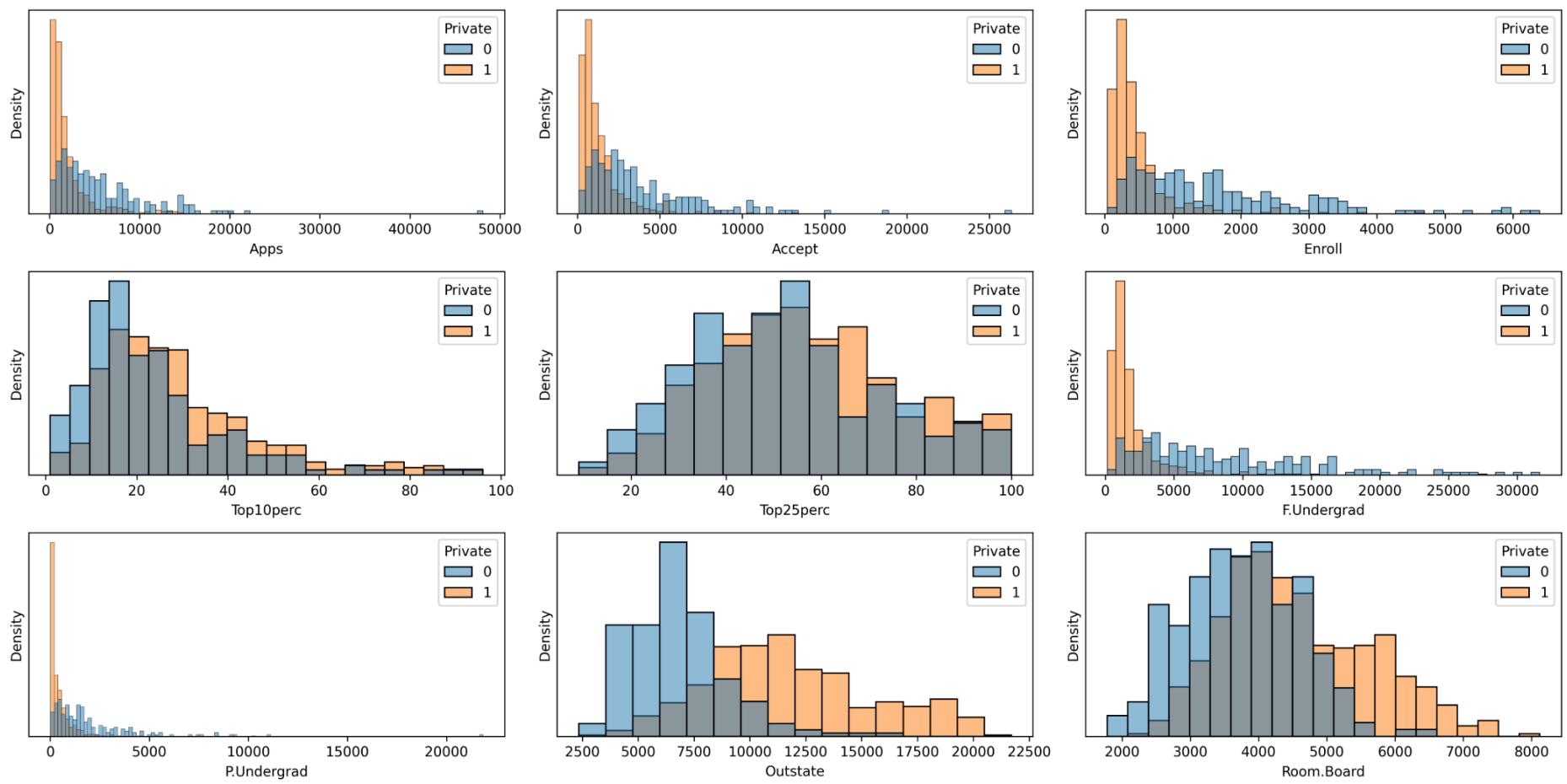
2.2 相关矩阵



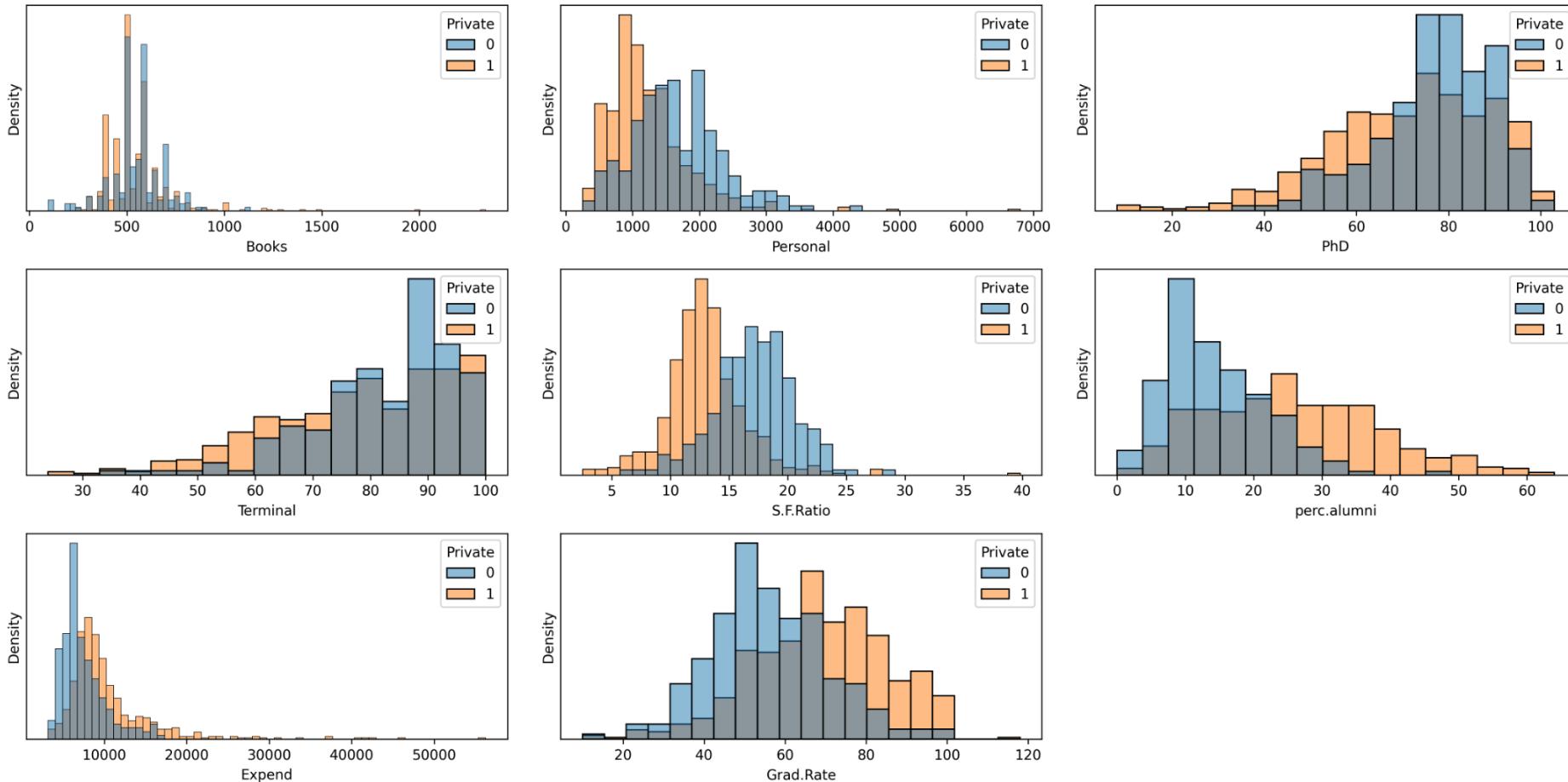
2.2 相关矩阵

1. 从整体来看，申请数 (Apps)、录取数 (Accept) 与入学数 (Enroll) 呈现显著的强正相关，显示出“招生漏斗”的现象
2. 教师中博士比例 (PhD) 与终极学位比例 (Terminal) 的相关系数较高
3. 外州学费 (Outstate) 与生均教学支出 (Expend) 的正相关则体现了资源投入与收费水平的正向关联，印证了高学费、高支出的精英大学的存在
4. 师生比 (S.F.Ratio) 与教师博士比例 (PhD) 呈现约一定的负相关性
5. 私立学校 (Private) 与师生比的负相关则表明私立学校倾向于维持更低的师生比
6. 书本费 (Books) 和个人开销 (Personal) 与其他特征的相关性较弱

2.3 公立学校和私立学校的部分特征数据分析



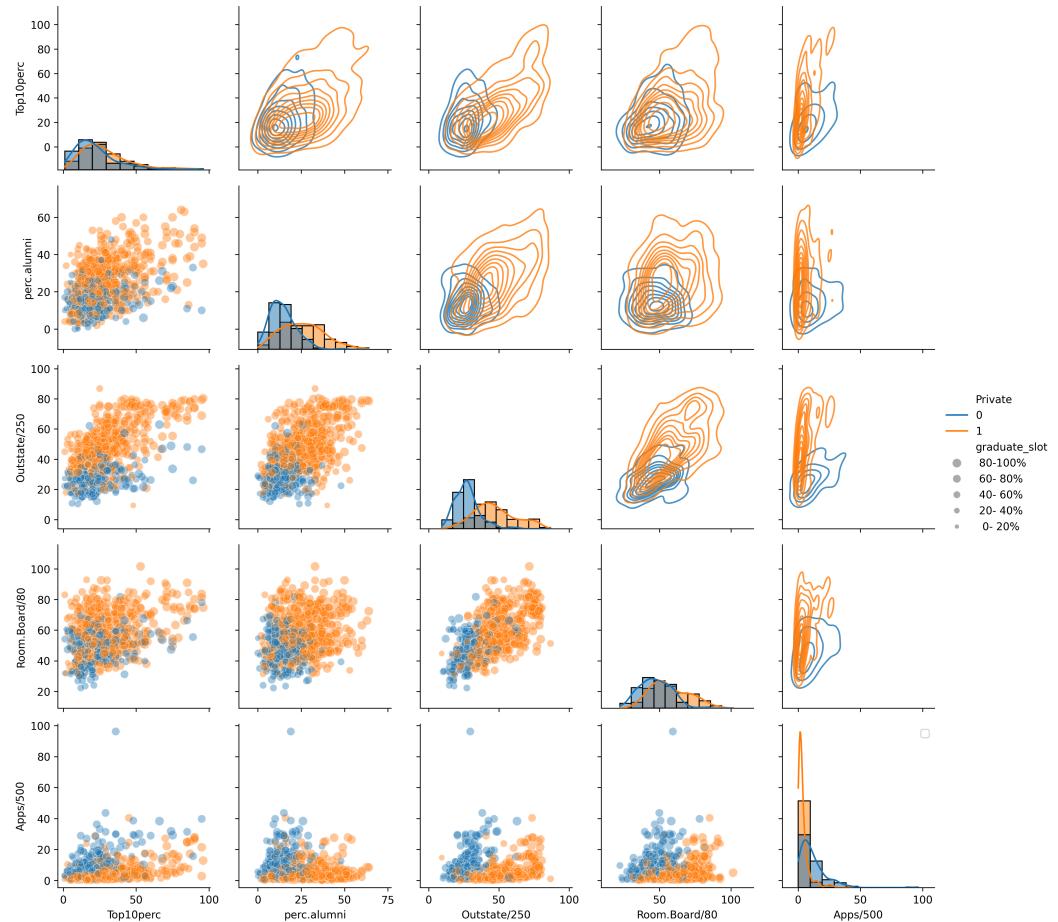
2.3 公立学校和私立学校的部分特征数据分析



2.3 公立学校和私立学校的部分特征数据分析

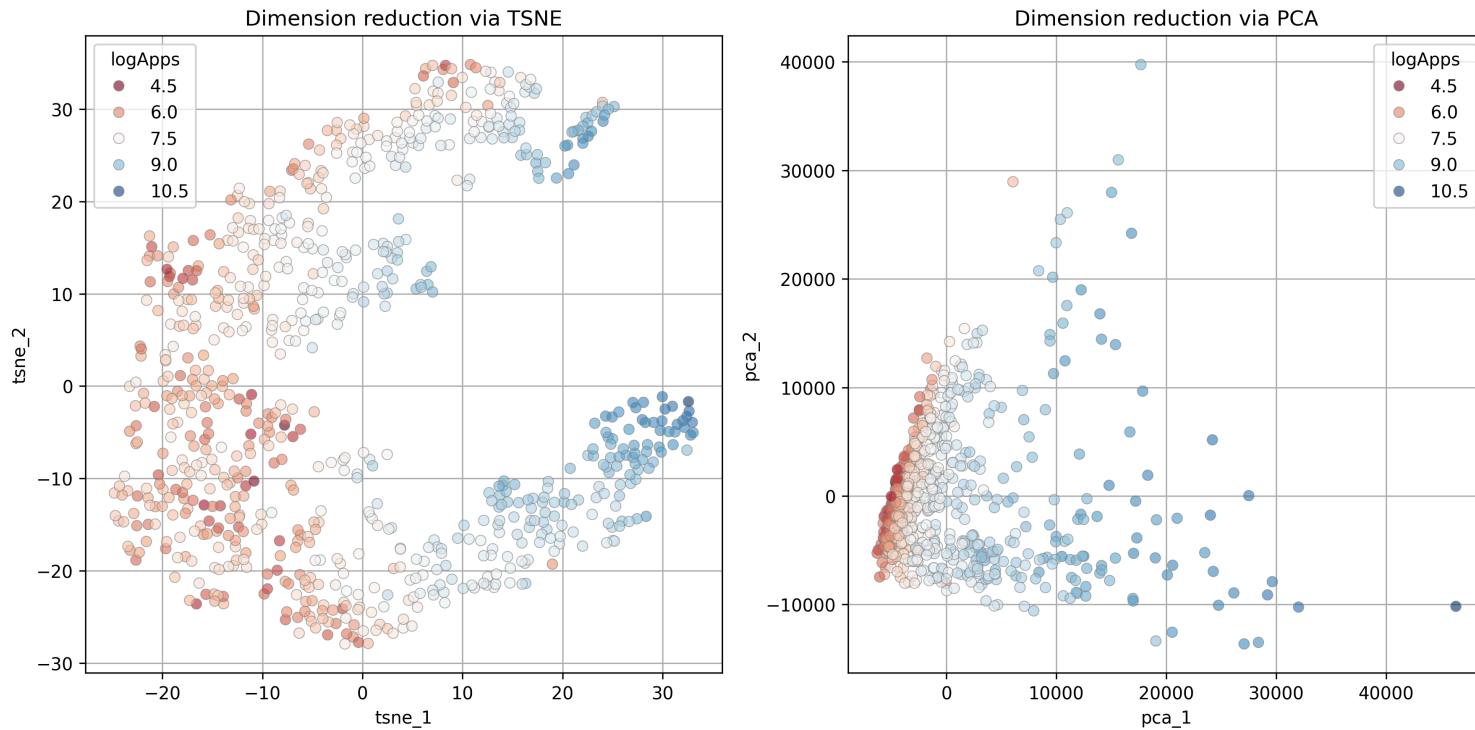
1. 公立学校和私立学校中来自中学中前 10% 或是前 25% 学生的占比分布、教师中终极学位比例、学校的生均开销和学生的书本开销差距不大
2. 申请相关指标（包括申请、录取、入学人数）、全日制和非全日制本科生人数中，私立学校都比公立学校对应的特征值少且分布集中
3. 私立学校学生的学费和食宿开支（Outstate 的约一万至两万美元和 Room.Board 的约三千美元至约七千美元）相比公立学校更高，毕业率（70% 左右）和捐款校友比例（约为 15% ~ 45%）也更高；而学生的预估个人开销则更低（约一千美元）。

2.3 公立学校和私立学校的部分特征数据分析



- 录取中学前十分之一的学生的学校校友捐赠比例更高、外州学生学费更高、学生们的生活开销也更高
- 对于公立学校的申请人数和前十百分之一学生比例、校友捐赠比例、外州学生学费和生活开销正相关
- 私立学校的申请人数则与这些指标的关系不大

2.4 降维可视化



- 高申请人数学校降维后的特征点大多位于图的右侧，低申请人数学校聚集于左侧
- 可以猜测，低申请人数的学校可能其特征更为相近，高申请人数的名校也许各有特色

2.5 正态性检验

我们对数据的每个特征列都做 K-S 检验，得到所有的 p 值都为零：这说明所有列都不遵从正态分布。

目录

1. 研究背景与目的	1
2. 探索性数据分析	3
2.1 重要统计量分析	5
2.2 相关矩阵	7
2.3 公立学校和私立学校的部分特征数据分析	9
2.4 降维可视化	13
2.5 正态性检验	14
3. 方法和实验	15
3.1 数据预处理	16
3.2 线性回归及其变体	17
3.3 决策树	24
4. 结论和讨论	27

3.1 数据预处理

通过小节 2.1 可以看到，数据中各域的数量级差距悬殊，这会导致在线性回归等训练过程中权重参数被压制接近于 0。于是可以统一将所有数据域放缩至 $[-1, 1]$ 这个区间。实现这个要求可以使用 `sklearn` 模块的 `MaxAbsScalar`，其原理为该数据域的数据统一除以最大的绝对值，即

$$x'_i \leftarrow \frac{x_i}{\max_i(x_i)}.$$

这样我们就得到了一系列数量级相同的归一化特征。归一化后对数据集进行的随机切分，为训练集占 70%，验证集占 30%。至于回归的目标，综合考虑先前分析得到的申请数量的分布，我们将申请人数取对数，得到对数申请人数 (`logApp`)，并将其作为回归目标。

3.2 线性回归及其变体

下面简要介绍所用的线性回归基线方法族。线性回归可以用直线拟合、向数据矩阵 \mathbf{X} 的列空间做正交投影、极大似然估计等视角进行理解和解释。记数据集为 $D = \{(x_{\{n\}}, y_n)\}_{n=1}^N$, 定义对标签的预测函数为下面的线性形式

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D = \mathbf{w}^\top \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix}$$

其中 $\mathbf{x} \in \mathbb{R}^D$, $\mathbf{w} \in \mathbb{R}^{D+1}$. 这是线性回归的基本形式。若将特征和对应的标签堆叠起来, 即

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

3.2 线性回归及其变体

则线性回归的预测结果为 $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\omega}$, 我们要求预测结果 $\hat{\mathbf{y}}$ 距离真实目标 \mathbf{y} 越近越好, 这就得到了线性回归的目标函数:

$$J(\boldsymbol{\omega}) = \frac{1}{2N} \sum_{n=1}^N (y_n - \hat{y}_n)^2 = \frac{1}{2N} \|\hat{\mathbf{y}} - \mathbf{y}\|_{[2^2]} = \frac{1}{2N} \|\mathbf{X}\boldsymbol{\omega} - \mathbf{y}\|_2^2,$$

这是一个拥有光滑凸目标函数的优化问题, 若 $\mathbf{X}^\top \mathbf{X}$ 可逆, 令 $J(\boldsymbol{\omega})$ 的梯度为零, 我们能得到其解析解

$$\boldsymbol{\omega}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

实际操作中, 我们会在目标 $J(\boldsymbol{\omega})$ 中加入正则项 $R(\boldsymbol{\omega})$, 通过对权重向量进行软限制的方式防止过拟合。

3.2 线性回归及其变体

修改过后的目标函数为

$$J(\omega) = \frac{1}{2N} \|X\omega - y\|_2^2 + \lambda R(\omega),$$

其中 $\lambda \geq 0$ 是需要人为给定的权重超参数。常用的正则项可以是权重向量的 L2 范数（这将得到岭回归）或 L1 范数（这将得到 LASSO 回归）。若选取的 L2 范数，问题总是存在整洁的解析解：

$$\omega^* = (X^\top X + \lambda I)^{-1} X^\top y.$$

从计算上看，这会使得括号中的矩阵可逆，此时该模型总是有解析解。

3.2 线性回归及其变体

ElasticNet 方法结合了岭回归和 LASSO 回归，其目标函数为

$$J(\omega) = \frac{1}{2N} \|X\omega - y\|_2^2 + \lambda [\alpha \|\omega\|_1 + (1 - \alpha) \|\omega\|_2^2].$$

我们使用标准线性回归、岭回归、LASSO 回归 和 ElasticNet 回归这四种方法为基准方法，使用其余特征预测对数申请人数。对于标准线性回归，允许其拟合截距项 (interception)；对岭回归，设置正则项权重为 $\alpha = 0.01$ ；对 LASSO 回归，设置正则项权重为 $\alpha = 0.001$ ，对 ElasticNet，设置正则项权重为 $\alpha = 0.001$ ，且设置 L1 和 L2 正则项的权重相同 (`l1_ratio=0.5`)。其他为 `sklearn` 中的默认参数，详见附录。

3.2 线性回归及其变体

模型	R^2	均方损失 (MSE)
LinearRegression	0.75	0.30
Ridge	0.75	0.30
Lasso	0.74	0.31
ElasticNet	0.74	0.31

可见四个基准模型的预测损失 MSE 在 0.3 左右, R^2 分数在 0.75 左右, 这说明, 一定程度上线性模型可以拟合数据的大致趋势。

3.2 线性回归及其变体

- 四种线性基线模型在训练集上学习得到的大部分权重相差不大，由于 LASSO 和 ElasticNet 中含有 L1 范数，这会导致学习得到的权重更加稀疏
- 模型对学校是否为私立学校、录取率、入学率、师生比例、生均开支和毕业率有大权重，其中录取率的权重为除去偏置项后最大，这一定程度上可以反应它和择校策略之间的联系
- 录取率越高、毕业率、生均开销、教师 PhD 占比、师生比例越高的学校，对应申请人数越多。

Models	Intercept	Private	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
LinearRegression	4.62	-0.67	3.49	2.74	0.26	0.41	-1.08	0.84	0.48	0.73	0.80	0.32	0.72	-0.01	1.18	-0.25	0.99	0.94
Ridge	4.62	-0.67	3.46	2.71	0.26	0.41	-1.03	0.82	0.49	0.73	0.80	0.31	0.72	-0.01	1.18	-0.25	0.98	0.94
Lasso	4.83	-0.66	3.19	2.06	0.26	0.39	-0.00	0.21	0.46	0.76	0.64	0.20	0.72	0.00	0.92	-0.21	0.73	0.87
ElasticNet	4.77	-0.67	2.80	2.21	0.26	0.41	0.00	0.37	0.51	0.77	0.67	0.22	0.72	0.00	0.95	-0.24	0.76	0.90

Fitted Params

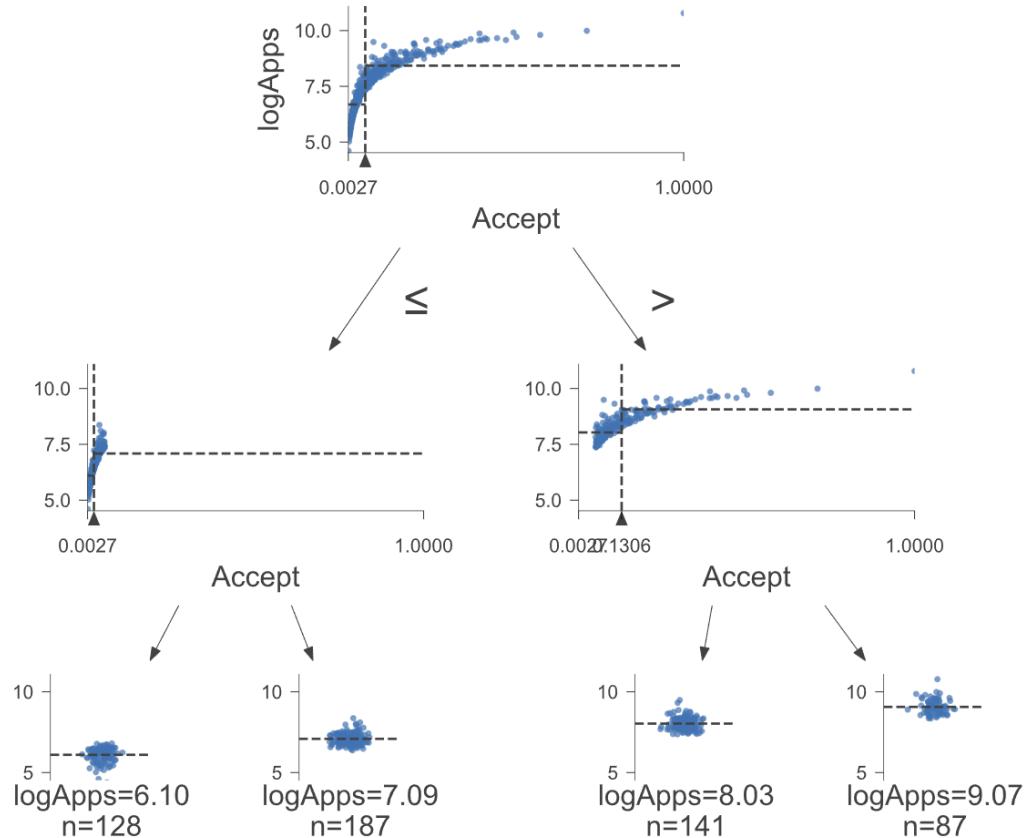
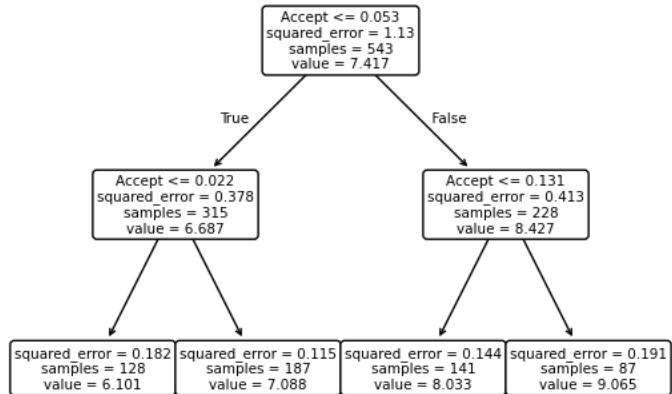
3.2 线性回归及其变体

如果将 Accept 和 Enroll 两列删去，得到的训练结果中，四个模型的 R^2 分数均为 0.68，MSE 分数均为 0.39。可以看到除了被删去的两列外，大体上贡献较大的特征和删除之前相同，即是否为私立、师生比、生均花费和毕业率。

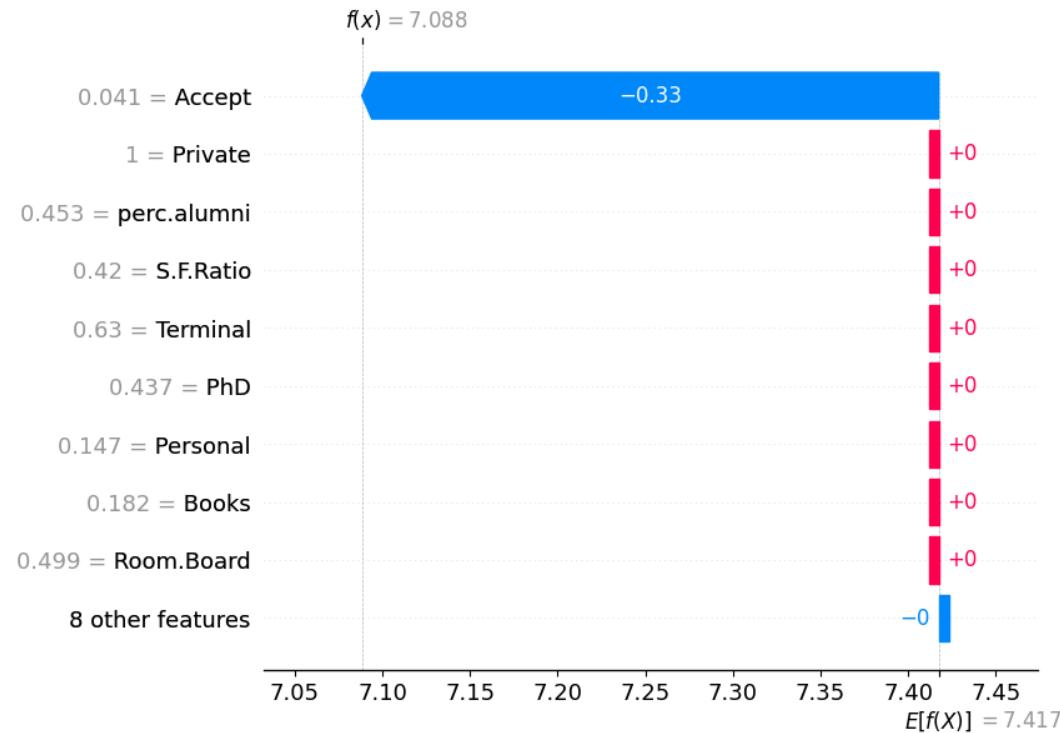
Models	Intercept	Private	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
LinearRegression	4.37	-0.65	0.00	0.58	3.79	0.07	0.67	0.93	0.91	0.11	0.85	-0.17	1.23	-0.35	1.25	1.12
Ridge	4.37	-0.65	0.01	0.58	3.79	0.07	0.67	0.93	0.90	0.11	0.84	-0.17	1.22	-0.35	1.25	1.12
Lasso	4.56	-0.65	0.05	0.55	3.82	0.00	0.62	0.91	0.73	0.01	0.73	-0.00	0.99	-0.31	0.99	1.06
ElasticNet	4.54	-0.68	0.10	0.54	3.66	0.05	0.66	0.91	0.74	0.09	0.74	-0.00	1.01	-0.34	0.96	1.07

3.3 决策树

最后考虑使用决策树算法对同样的数据集进行预测。和分类决策树类似，决策树对节点中的数据按照最优划分条件递归划分，直至满足停机条件。回归决策树的预测方法为将叶节点对应的训练集的目标值的平均作为该叶节点的预测目标值。我们同样采用 sklearn 中的 `DecisionTreeRegressor` 类进行训练，并约束其分裂指标为平方损失 (`squared_error`)、最大深度为 2、最小可分裂数据量为 10，以保证模型的泛化性能。该模型在测试集上达到 0.85 的 R^2 分数和 0.178 的 MSE 分数，相比于线性回归基准模型低，效果较回归模型更好。



3.3 决策树



如果像线性模型那样去除相关的两列，并使用 optuna 模块搜索最优参数，得到的模型在测试集上可以达到 0.28 的 MSE 值，此时 **F.Undergrad** 拥有最高的 Shap 值。

目录

1. 研究背景与目的	1
2. 探索性数据分析	3
2.1 重要统计量分析	5
2.2 相关矩阵	7
2.3 公立学校和私立学校的部分特征数据分析	9
2.4 降维可视化	13
2.5 正态性检验	14
3. 方法和实验	15
3.1 数据预处理	16
3.2 线性回归及其变体	17
3.3 决策树	24
4. 结论和讨论	27

4. 结论和讨论

通过数据分析和建模预测可以看出，各个学校的申请人数和学校的实力、师资等等有重要关系。申请者同时也关注在学校的开销、学校倾注在学生身上的花费，以及师生比，这可以看出学校对学生的重视程度。申请者还看重毕业率，这预示着他们将来成功从这所大学毕业拿到文凭的概率。

由于本数据的记录时间很早，到现在已有近三十年，上述结果可能不足以说明当今世界的高校申请形势。同时在分析中遇到的诸多疑似“倒果为因”的现象，使对相关数据的因果联系的建模与发现变得必要。

Thanks for Listening :)