升学数据分析报告

摘要：本案例通过对 UCLA 的公开申请数据做描述性分析，初步判断申请几率与各潜在影响因素的关联。结果显示，英语成绩，在校表现和综合实力都与最终录取的机率有关。

一、 背景介绍

目前全球教育实力较强的院校主要集中在美国，欧洲等地，因此许多中国学生选择出国以获得更好的教育资源。截至 2021 年我国留学市场规模已达到 6000 亿元，年出国人数超过 71 万人，并且还在持续扩大。在所有出国的学生中 60%以上是本科及以上留学，说明出国读研是行业内主要的细分市场。

目前我国留学行业还存在一些问题。宏观上平台提供的服务体系性不强，服务质量非常依赖个人能力。微观上研发设计能力不足，服务无法匹配用户的特点，满足不了个性化的需求。本案例的研究意义是从数据入手，帮助平台了解申请成功率的影响因素，并据此形成兼顾体系性与个性化的留学服务定制产品。

二、 数据描述

本案例使用 UCLA 在 2019 年的申请数据，共 500 条。数据来自数据竞赛网站 Kaggle，共包含 8 个变量，其中因变量一个（录取几率），自变量包括英语成绩两个（GRE 成绩，TOEFL 成绩），在校表现两个（学校排名和在校 GPA），综合实力三个（自我陈述，推荐信和研究经历得分）。详细的变量说明如表 1 所示。

表 1 数据变量说明表

变量类型		变量名称	详细说明	取值范围
因变量		录取几率	连续变量	0-1
自变量	英语成绩	GRE 成绩	单位：分	0-340
		TOEFL 成绩	单位：分	0-120
	在校表现	学校排名	定性变量（5 水平）	1-5
		在校 GPA	连续变量	0-10
		自我陈述得	定性变量	1-5

	综合实力	分	（5 水平）	
		推荐信得分	定性变量（5 水平）	1-5
		研究经历	定性变量（2 水平）	是/否

三、 描述分析

首先对数据进行描述性分析，初步判断申请成功率与各潜在影响因素的关联。

（一）因变量

从图 1 中可以发现：大部分申请者的录取几率超过了 50%，数据主要分布在 0.6-0.8 之间，而当录取几率低于 0.65 时，成功率越低申请者数量越少。这两个现象说明在申请 UCLA 时大部分申请者都做了不少准备，因此这所学校的竞争相对激烈，内卷情况比较严重。录取几率在 0.8-0.9 的区间内时，申请者数量相差不多。这说明最有优势的那批申请者分布比较均匀。

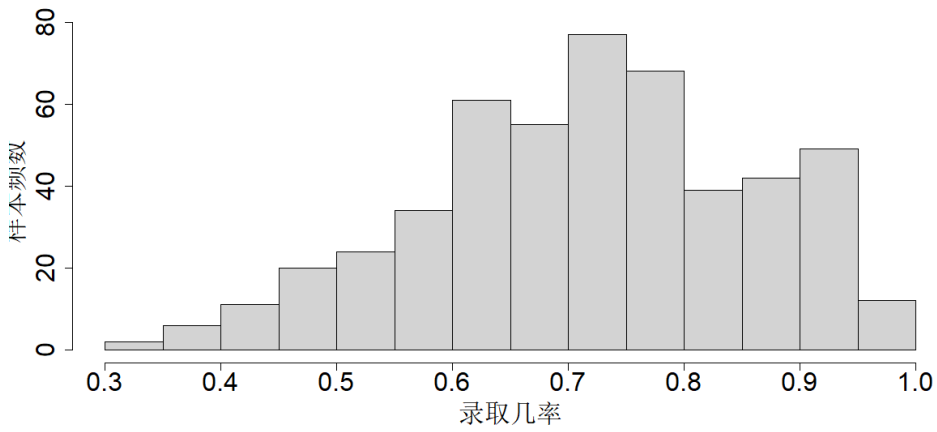

图 1 录取几率直方图

（二）自变量

从图 2 中可以看出 GRE 和 TOEFL 成绩和录取几率都呈很强的正相关性，计算

相关系数分别得到 0.81 和 0.79。GRE 与 TOEFL 是目前西方国家最权威的英语标准化考试，分数直观地反映了学生的英语能力，而能力越高的学生适应国外全英语教学的速度越快，因此更加优秀。

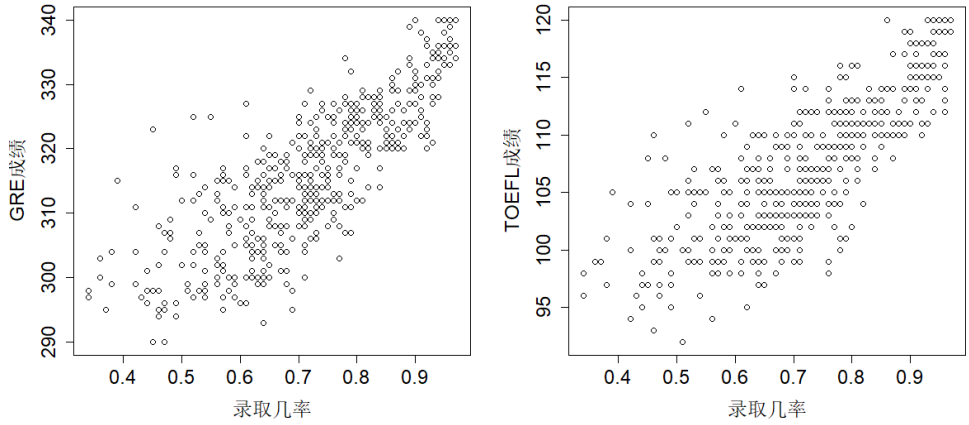

图 2 录取几率与英语成绩散点图

从图 3 中可以看出，显然学校排名与录取几率有很强的正相关性。排名越高平均值越高，最好的第五档学校录取几率的方差远小于其他 4 个档次，这也许是因为学校出色的学生在各方面都会比较优秀，导致整体被拒绝的可能性不大。

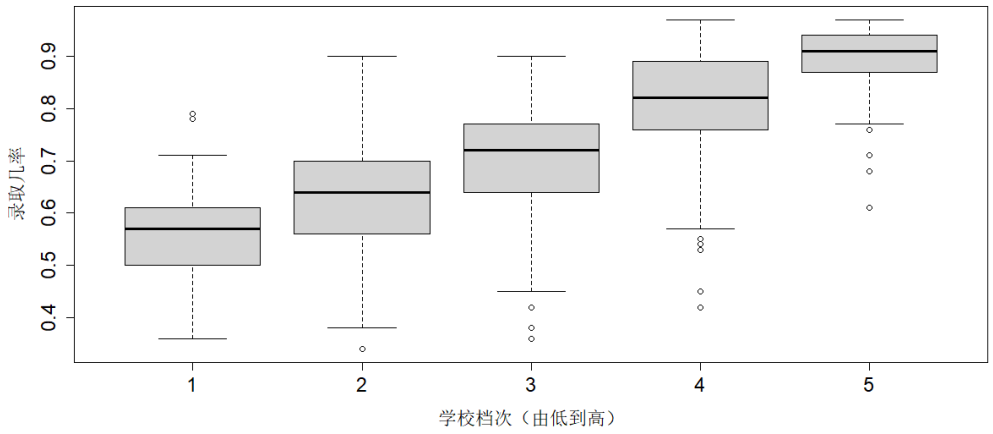

图 3 录取几率与学校排名箱线图

通过图 4 可以发现自我陈述和推荐信得分越高，平均录取几率越大，有研究经历的申请者平均录取几率要高于没有经历的人。自我陈述和推荐信得分最高和最低的学生学生录取率的方差总体上最小。背后的原因可能是在所有申请者中最优秀和最差的学生一般都比较突出，面试官很轻松就能做出是否录取的决定。

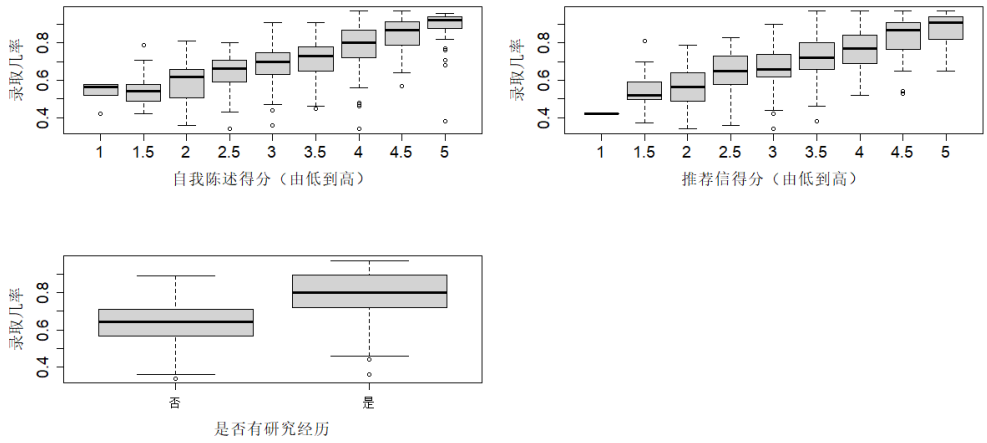

图 4 综合实力与录取几率箱线图

四、 小结

英语成绩，包括 GRE 成绩和 TOEFL 成绩都和录取几率有高相关性。学校排名，自我陈述和推荐信得分越高，录取几率越大。有研究经历的人比没有的申请者更有优势。