留学申请竞争力影响因素探究

——基于多种回归模型的数据分析

内容提要

留学申请竞争愈加激烈，如何提高申请者的竞争力是值得研究的课题。我们基于多种回归模型，包括普通最小二乘回归模型、AIC、BIC 选模型、LASSO 回归模型、回归树模型、加权最小二乘回归模型、分位数回归模型，分析留学申请数据，对留学申请者给出了一定的竞争力提升策略，最后对上述模型进行了思考与总结，从统计方法的整体视角给出了我们对不同模型的理解。

报告撰写人：孙翰澈、油天宇、白启东

目录

1.背景介绍

新冠疫情爆发以前，教育部公开数据显示，2013 年至 2019 年我国出国留学的学生数量呈逐年递增的趋势，在 2019 年达到 70.35 万人。受到疫情冲击之后，2020 年留学人数锐减，在 2021 年回暖。大体上，留学升温迹象十分明显，将来一段时间内仍然会延续历史趋势。

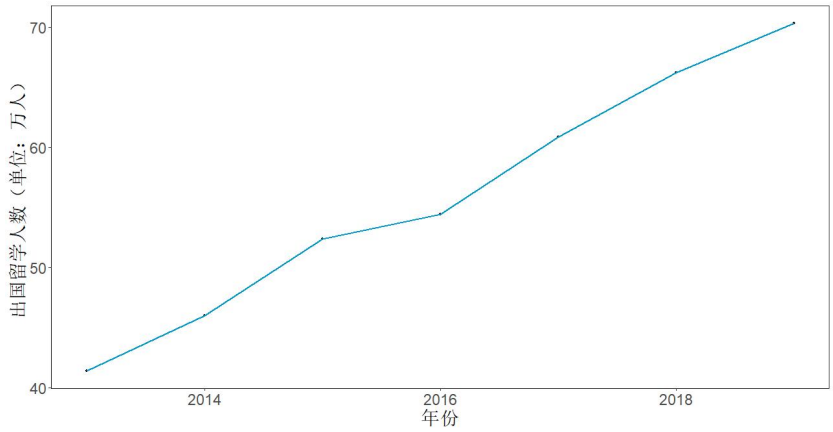

图 1 2013 年至 2019 年我国出国留学的学生数量

除了开阔国际视野、感受异国文化、提高求职竞争力等个人追求，以及相较于考研，留学申请机会与选择更多之外，促进留学不断升温的还有社会客观因素的变化，主要有：

一方面，国内政策利好出国留学。我国针对留学回国人才出台了一系列政策，包括就业服务，例如 2022 年 9 月 17 日上线的国家留学人才就业平台；搭建创新创业平台，例如举办“春晖杯”中国留学人员创新创业大赛；落户政策，以上海为例，世界排名前 50 高校留学生可直接落户；免税购车等。同时，疫情防控措施的变化也为出国留学提供了便利：“二十条”和“新十条”出台，减少了境外人员入境隔离时长，同时取消了航班熔断机制，大幅降低了留学生回国的时间成本与金钱成本。

另一方面，目前国内升学、就业压力较大。2023 年全国硕士研究生统一招生考试报名人数高达 548.4 万人，相比于去年增加 91.4 万人，达到近十年来的考研人数顶峰。有限的研究生名额与涨势凶猛的报名人数，带来的是更加白热化的竞争。另外，青年失业率仍然较高，同时结构性矛盾为就业市场带来了一定压力。出国留学是避开国内激烈竞争的选择之一。

根据新东方《2022 年留学申请白皮书》可以发现，我国出国留学的群体中绝大多数是本科毕业生，专注于研究本科毕业生的留学申请情况是有代表性与参考价值的。研究生阶段的留学申请所需要的材料包括标化成绩，例如 TOEFL、IELTS、GRE、GMAT 等；在校成绩，主要是 CGPA；其余材料，主要是推荐信、自我陈述、本科学校排名等。《白皮书》还显示，我国出国留学的人数逐年增长，竞争也变得激烈起来，不少学校对申请者的各项要求都或明或暗地提高了标准。这也使得很多申请者在留学申请时找不到努力的方向，产生了许多问题，例如是否应该参加科研项目？还需不需要参加竞赛丰富自己的履历？GPA 是不是留学申请中最重要的？

2.研究目的

我们的项目旨在基于留学申请数据，通过回归分析的方法，找出影响最终申请结果的因素，最终解决留学申请者的困惑，为其应该针对性地提升自己的哪些方面提供参考意见。同时，我们还希望使用多种回归模型，来探索不同模型的应用场景及其预测效果。

3.数据来源和说明

我们使用的数据集来源于 Kaggle，包括 400 条数据，以及 9 项指标，其具体情况如表 1 所示：

表 1 数据说明

变量	符号	类别	说明
序号	Serial No.	离散型	取值为 1 至 400 之间的整数
GRE 得分	GRE Score	连续型	可能取值在 0 至 340 之间
TOEFL 得分	TOEFL Score	连续型	可能取值在 0 至 120 之间
本科院校评分	University Rating	离散型	取值为 1 至 5 之间的整数
目的陈述评分	SOP	离散型	取值在 1 至 5 之间，步长为 0.5
推荐信评分	LOR	离散型	取值在 1 至 5 之间，步长为 0.5
累计平均学分绩点	CGPA	连续型	可能取值在 0 至 10 之间
科研经历	Research	离散型	取值为 0 或 1
录取几率	Chance of Admit	连续型	取值在 0 至 1 之间

该数据集涵盖了申请出国时，目标院校将会参考的几乎所有指标。我们选取

录取率为因变量，除序号外其余各变量为自变量。值得注意的是，录取几率为 0 至 1 之间的连续型数值，该指标反映的是在同水平下的所有申请者总体中，留学申请成功的学生数量占比。此外，CGPA 为 10 分制，与一般 4 分制 CGPA 不同，但其计算规则大体类似，并不影响后续的数据分析。

4.探索性数据分析

4.1 数据预处理

4.1.1 缺失值处理

经判断，该数据集没有缺失值，无需对其进行缺失值处理。

4.1.2 新变量定义

在表 1 中可以发现，SOP 与 LOR 的评分均为 9 类，分类较多导致每类样本较少，可能影响后续模型的效果；同时这两类评分均为专家打分，具有主观性，不同的水平对应的评分应该是一个取值范围，而非一个精确的数值。因此，我们定义了目的陈述档次 SOP1，如下述公式所示：

$$SOP1 = \begin{cases} 1, & 1 \le SOP \le 2 \\ 2, & 2 < SOP < 4 \\ 3, & 4 \le SOP \le 5 \end{cases}$$

SOP1 取值 1、2、3 分别代表该申请者的目的陈述文案一般、良好、优秀。推荐信档次 LOR1 类比 SOP1 定义即可。

4.1.3 划分预测集

后续过程中需要评价模型的预测效果，于是我们将该数据集划分为训练集与预测集。由于数据集规模较小，我们选择原数据集的 20%作为预测集，即预测集中数据量为 80。我们采取如下算法进行预测集划分：

计算录取几率的极差R，令$d = \frac{R}{5}$。以d为步长将录取几率划分为 5 个区间，分别统计 5 个区间内录取几率的数量。采用分层抽样法，随机抽取每个区间的数

据，总共抽取 80 个作为预测集。

这样的划分方法可以保证预测集与原数据集的分布大致类似。另外，训练集即预测集相对于原数据集的补集。

4.2 描述性统计

我们对原数据集进行了下列描述性统计，来初步观察数据的面貌。

对连续型变量计算最小值、中位数、均值、最大值、标准差，结果如表 2 所示：

表 2 连续型变量的数字特征

	最小值	中位数	均值	最大值	标准差
GRE Score	290.000	317.000	316.808	340.000	11.474
TOEFL Score	92.000	107.000	107.410	120.000	6.070
CGPA	6.800	8.610	8.599	9.920	0.596
Chance of Admit	0.340	0.730	0.724	0.970	0.143

表 2 反映了数据集中留学申请者的“硬实力”水平。GRE、CGPA 与录取几率呈左偏分布，表明有部分 GRE、CGPA 水平，以及录取几率较不理想的申请者。TOEFL 成绩呈现右偏分布，表明有部分 TOEFL 水平相对优秀的申请者。观察这四个变量均值与中位数的具体数值可知，这部分申请者的总体水平是比较优秀的。

我们以录取几率为纵坐标，对连续型自变量绘制了如图 2 所示的散点图：

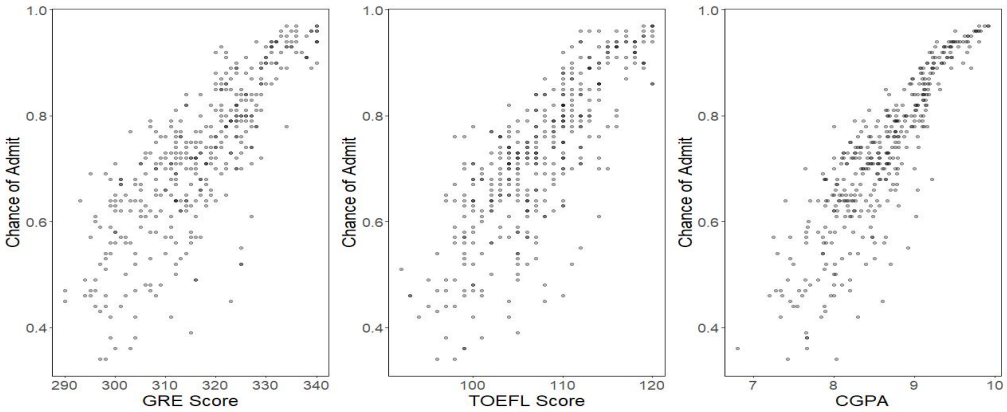

图 2 连续型自变量散点图

由图 2 可知，随着 GRE、TOEFL、CGPA 的增长，录取几率也相应提高，这符合我们的常识。并且三者均与录取几率显示出线性关系，这也为我们后续采用线性模型进行回归分析提供了依据。

同样地，对离散型自变量绘制如图 3 所示箱线图，同时加入散点，以便观察

数据的集中趋势：

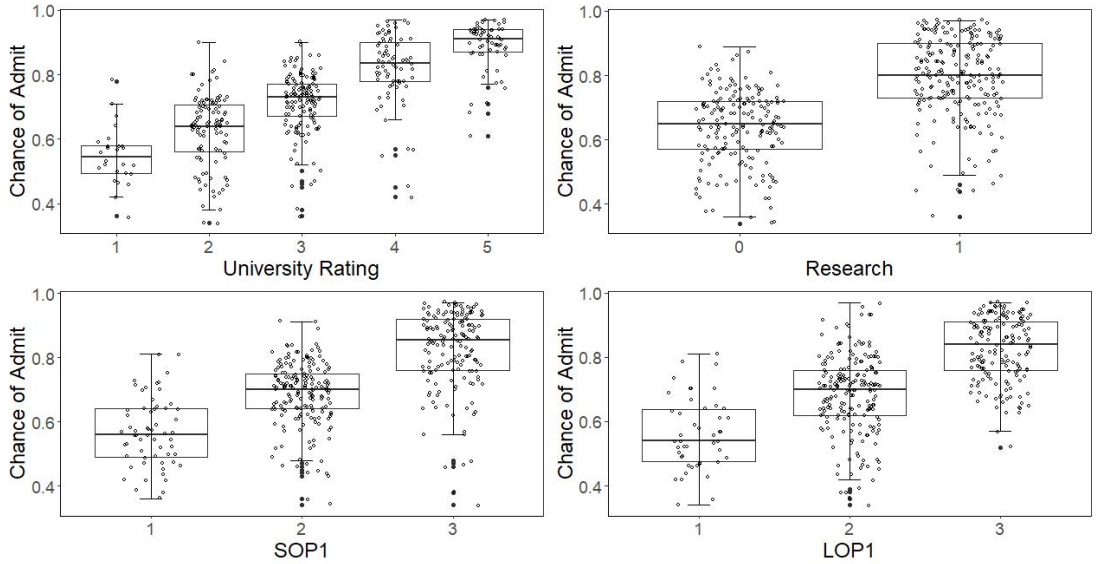

图 3 离散型自变量箱线图

从图 3 中可以发现，本科院校评分越高，录取几率越大。但也不乏有本科院校评分较低，录取几率大的申请者。整体上，有科研经历要好于没有科研经历。此外，目的陈述与推荐信的档次越高，录取几率越大。

对于录取几率，我们绘制了如图 4 所示直方图，可以直观地看出录取几率的分布情况：

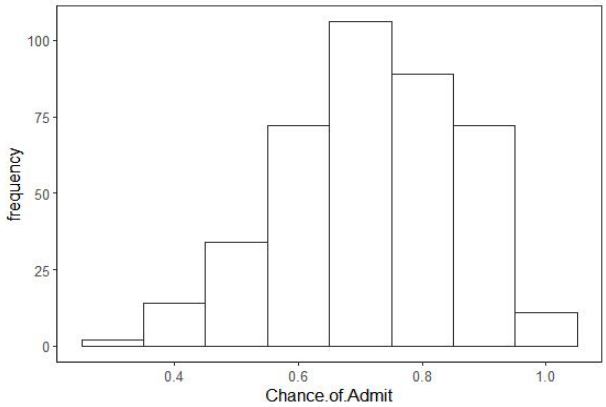

图 4 录取几率直方图

最后，我们绘制了如图 5 所示的所有连续型变量的相关系数矩阵热力图：

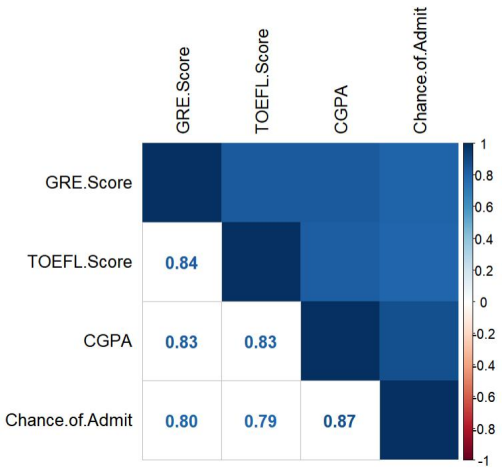

图 5 相关系数矩阵热力图

从图 5 中可以看出，这四个变量两两之间的相关性基本超过 0.8，因此在后续建模过程中，需要注意消除变量之间的复共线性。

5.数据建模

5.1 原始全模型

出于试探目的，我们对训练集直接进行普通最小二乘回归，基于此寻找自变量(GRE Score,TOEFL Score,University Rating,SOP1,LOR1,CGPA,Research)与因变量(Chance of Admit)之间的关系。将分类变量转换为因子变量后，理论上可以得到形如表 3 所示的线性回归模型（记作原始全模型）：

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{12} X_{12}$$

表 3 原始全模型

	Estimate	Std.Error	t-value	Pr(>|t|)	
(Intercept)	-1.231	0.144	-8.538	0.000	***
GRE.Score	0.002	0.001	3.027	0.003	**
TOEFL.Score	0.003	0.001	2.219	0.027	*
University.Rating2	-0.013	0.017	-0.776	0.438	
University.Rating3	-0.012	0.019	-0.617	0.538	
University.Rating4	-0.003	0.022	-0.122	0.903	
University.Rating5	0.022	0.024	0.888	0.375	
SOP12	0.025	0.013	1.993	0.047	*
SOP13	0.009	0.015	0.599	0.550	
LOR12	0.012	0.014	0.881	0.379	
LOR13	0.042	0.016	2.593	0.010	**
CGPA	0.115	0.013	8.922	0.000	***

Research1	0.030	0.009	3.389	0.001	***
RSE	0.063		F: p-value	0.000	
R-squared	0.810		Adjust R^2	0.802	

模型显示，在 0.05 的置信水平下，GRE、TOEFL、CGPA、是否有科研、中档的目的陈述、优秀的推荐信均与录取几率显著呈现正相关。

对该模型进行必要的回归诊断，来检验基于最小二乘的估计方法所需要的模型假设是否近似成立，以及判断数据中有没有异常值或强影响点。绘制如图 6 所示的四张图：

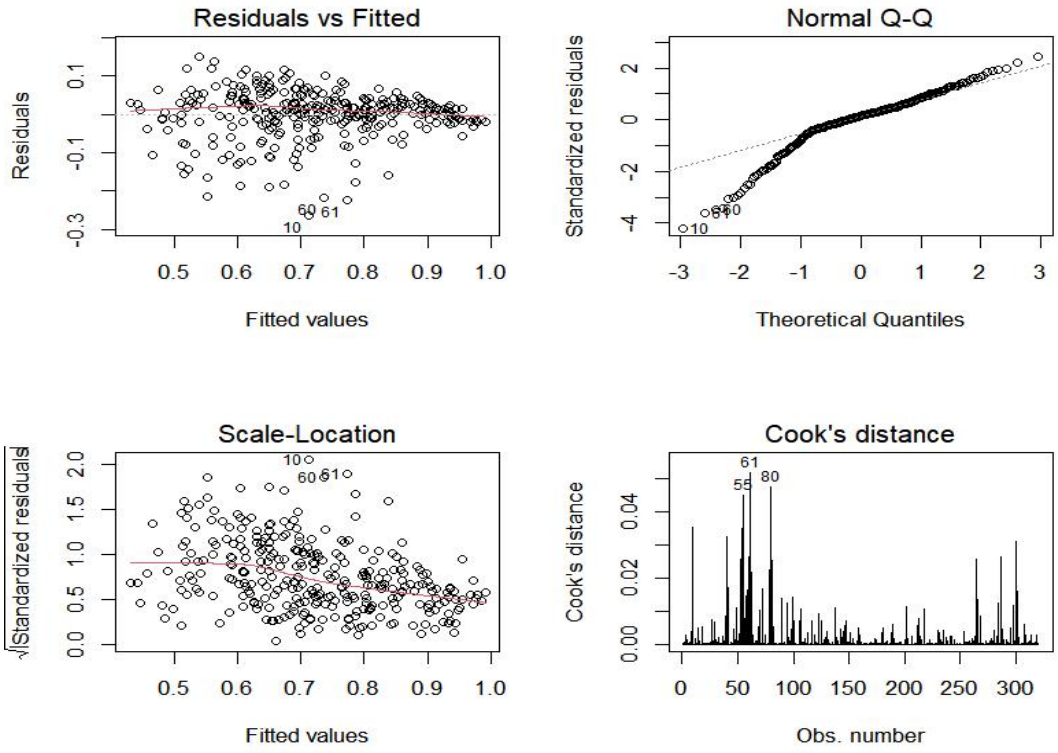

图 6 原始全模型回归诊断

观察图 6 左上角残差图，可以发现数据点从左向右逐渐集中，呈现向左开口的喇叭状，说明残差并不满足同方差的假设条件。另外，观察图 6 右上角 Q-Q 图，可以看出有较多数据点不在同一直线上，且有明显转折，说明随机误差不满足正态性的假设条件，因此可以得出结论：未经任何处理的数据并不满足 Gauss-Markov 条件以及一般的正态性假定。因此我们不对其回归系数等模型具体参数进行分析。

5.2 Box-Cox 变换

为了处理“异方差，非正态”的问题我们采用 Box-Cox 变换。Box-Cox 变换是 Box 和 Cox 在 1964 年提出的一种广义幂变换方法，是统计建模中常用的一种数据变换，用于连续的响应变量不满足正态分布的情况。Box-Cox 变换之后，可以一定程度上解决上述问题。通过引入参数λ，Box-Cox 变换对回归因变量作如下变换。

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ lnY, & \lambda = 0 \end{cases}$$

λ是一个待定的变换参数，为了确定λ的值，绘制如图 7 所示随λ变化时，对数似然函数值相应变化的曲线：

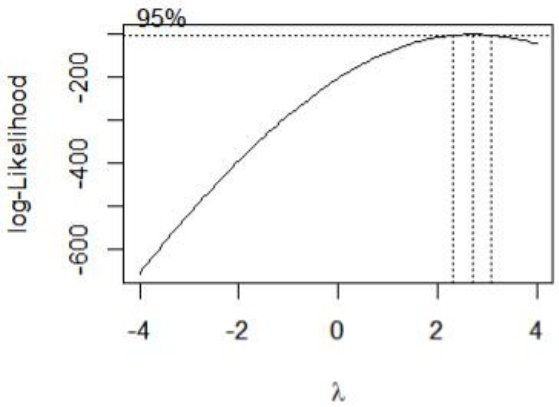

图 7 不同 λ 对应的对数似然函数值

计算得到对数似然函数在λ=2.7 时取得极大值。根据上述分析，我们对数据进行参数λ=2.7 的 Box-Cox 变换以后，再建立多类模型。

5.2.1 全模型

建立 Box-Cox 变换以后的全模型，重新进行回归诊断：

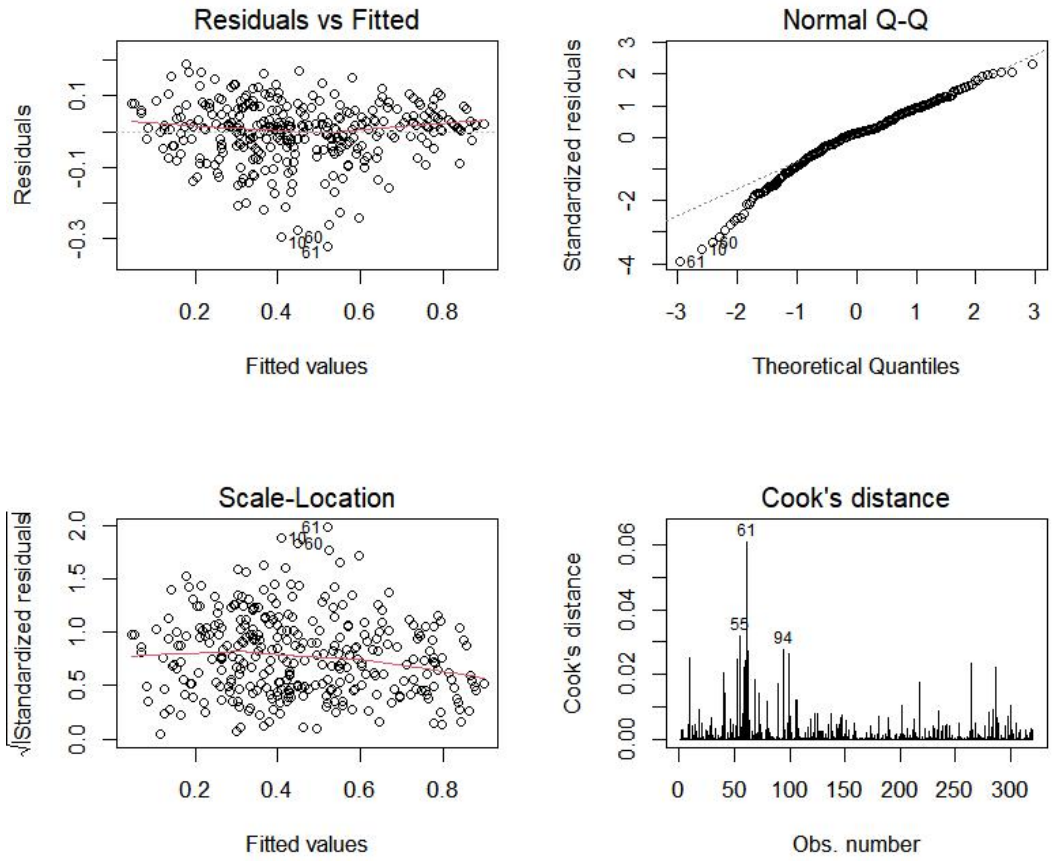

图 8 全模型回归诊断

与进行 Box-Cox 变换之前相比，残差分布呈现喇叭状的情况得到一定好转，并且在 Q-Q 图中，更少的点偏离了直线，且其明显转折的走势也得到改善。Box-Cox 变换后的数据基本满足线性回归模型的经典假设。此外，我们注意到点 61 的 Cook 距离较大，将其删除后利用剩余数据重新建模，结果如表 4 所示：

表 4 全模型

	Estimate	Std.Error	t-value	Pr(>|t|)	
(Intercept)	-2.505	0.191	-13.112	0.000	***
GRE.Score	0.003	0.001	3.424	0.001	***
TOEFL.Score	0.005	0.002	2.941	0.004	**
University.Rating2	-0.031	0.023	-1.347	0.179	
University.Rating3	-0.038	0.025	-1.534	0.126	
University.Rating4	-0.009	0.029	-0.295	0.768	
University.Rating5	0.058	0.032	1.813	0.071	.
SOP12	0.012	0.017	0.733	0.464	
SOP13	0.007	0.020	0.361	0.718	
LOR12	-0.003	0.018	-0.171	0.864	
LOR13	0.035	0.021	1.628	0.105	
CGPA	0.172	0.017	10.050	0.000	***

Research1	0.051	0.012	4.320	0.000	***
RSE	0.084		F: p-value	0.000	
R^2	0.857		Adjust R^2	0.850	

表 4 中，本科院校评分、SOP1 与 LOR1 被作为虚拟变量处理。可以看到，模型 F 检验的 P 值非常小，表明该模型是显著的，即自变量和因变量之间确实存在一定的关系，未调整与调整后的判决系数均较高，表明该模型对自变量和因变量之间的关系有较好的解释能力。容易发现， GRE、TOEFL、CGPA 与录取几率显著呈现正相关。若申请者有科研，其竞争力将会更加强劲。而本科院校是否为最高档次对其也有一定的正向影响，但仅在显著性水平 0.1 的情况下显著。

5.2.2 选模型

从以上对全模型的分析容易发现，有四项指标非常重要，一项指标较为重要，但是我们不能排除其他变量也有预测能力的可能，同时考虑模型的共线性问题，我们采用两种最常见的模型变量选择方法，即 AIC 和 BIC 来进行逐步回归，进而达成变量选择的目的。

（1）AIC

如果使用 AIC 来选择模型，可以得到如表 5 所示的模型估计结果：

表 5 AIC

	Estimate	**Std.Error**	**t-value**	**Pr(>|t|)**	
(Intercept)	-2.509	0.189	-13.265	0.000	***
GRE.Score	0.003	0.001	3.385	0.001	***
TOEFL.Score	0.005	0.002	3.063	0.002	**
University.Rating2	-0.026	0.022	-1.199	0.231	
University.Rating3	-0.033	0.024	-1.39	0.166	
University.Rating4	-0.006	0.028	-0.222	0.824	
University.Rating5	0.060	0.031	1.917	0.056	.
LOR12	-0.002	0.018	-0.111	0.912	
LOR13	0.035	0.021	1.702	0.090	.
CGPA	0.173	0.017	10.25	0.000	***
Research1	0.052	0.012	4.431	0.000	***
RSE	0.082		F: p-value	0.000	
R^2	0.863		Adjust R^2	0.858	

从表 5 中可以看出，AIC 认为除了全模型中较为重要的变量以外，LOR1 在 0.1 的显著性水平下对录取几率也有显著的影响。并且 AIC 模型的 RSE 略低于全模型，判决系数（R-square）略大于全模型。

（2）BIC

如果用 BIC 来选择模型，我们可以得到如表 6 所示的模型估计结果：

表 6 BIC

	Estimate	Std.Error	t-value	Pr(>|t|)	
(Intercept)	-2.549	0.190	-13.404	0.000	***
GRE.Score	0.003	0.001	3.351	0.001	***
TOEFL.Score	0.004	0.002	2.801	0.005	**
University.Rating2	-0.025	0.021	-1.189	0.235	
University.Rating3	-0.028	0.023	-1.257	0.210	
University.Rating4	0.010	0.027	0.362	0.718	
University.Rating5	0.078	0.030	2.568	0.011	*
CGPA	0.182	0.017	11.022	0.000	***
Research1	0.055	0.012	4.648	0.000	***
RSE	0.083		F: p-value	0.000	
R^2	0.859		adjust-R^2	0.855	

从表 6 中可以看出，BIC 与全模型所认为的重要变量相同，但 BIC 认为本科院校是否为最高档在 0.05 的显著性水平下对录取几率有显著影响。同时，其并不认为 LOR1 是重要的变量。此外，BIC 模型的判决系数 R-square 略低于 AIC，但略高于全模型。

5.2.3 LASSO 回归

LASSO 回归通过构造惩罚函数得到一个较为精炼的模型，它压缩一些回归系数，强制系数的绝对值之和小于某个固定值；同时设定一些回归系数为零。因此它保留了子集收缩的优点，是一种处理具有复共线性数据的有偏估计，同样适用于当前的分析场景。

完成 LASSO 回归建模需要选择参数λ的值，算法通过对不同的λ进行拟合，得到如表 7 所示结果：

表 7 确定 LASSO 回归参数 lambda

No.	Df	%Dev	Lambda
1	0	0.00	0.193700
2	1	13.48	0.176500
3	1	24.68	0.160800
…	…	…	…
52	6	84.84	0.001685
53	6	84.84	0.001535
54	6	84.84	0.001399

55	6	84.84	0.001275
56	6	84.84	0.001161
57	6	84.85	0.001058

其中，%Dev 表示模型的可解释偏差，值越大表明该模型包括了越多样本的信息，当计算进行到 57 次时，模型可解释偏差增幅不再明显，算法收敛，因此选取λ=0.001058。此时，LASSO 回归结果如表 8 所示：

表 8 LASSO 回归系数

	Coef
(Intercept)	-2.691
GRE.Score	0.003
TOEFL.Score	0.004
University.Rating	0.023
CGPA	0.174
Research	0.046
LOR1	0.014

可以发现，LASSO 回归对变量选择以后，剔除了变量 SOP1，认为其余自变量是较为重要的。

5.3 机器学习方法——回归树

为解决“异方差、非正态”的问题，我们尝试使用机器学习中的回归树方法，这是因为回归树是决策树的一种，它是非参数的方法，不需要预先对数据作任何概率分布假设。对数据采用该方法建模，并可视化回归树得到如图 9 所示的结果：

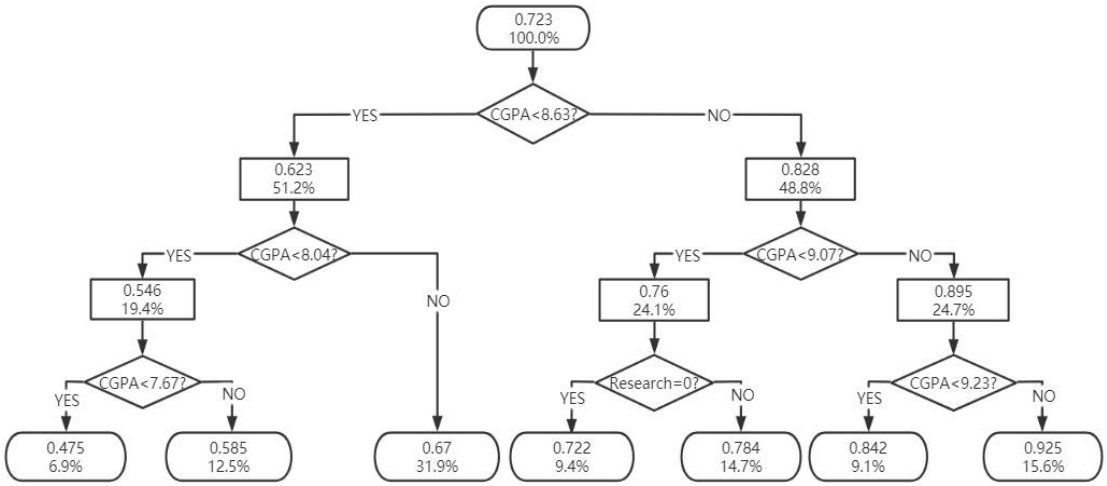

图 9 可视化回归树

可以发现，回归树将 CGPA 与有无科研经历作为唯二的分类依据，强调了 CGPA 与科研经历的重要性。

5.4 加权最小二乘回归

基于残差图显示的原始全模型的异方差性，也可考虑采用加权最小二乘法对原始全模型进行改进。传统的方法是选择与普通残差的等级相关系数最大的自变量构造权函数，且权函数通常取其幂函数，即：

$$W = x_j^m (j = 1,2,...,p)$$

于是首先计算三个连续性自变量与原始全模型残差绝对值的 spearman 等级相关系数，结果如表 9 所示：

表 9 spearman 等级相关系数表

	GRE.Score	TOEFL.Score	CGPA
Spearman correlation	-3.415E-01	-2.811E-01	-3.366E-01
p-value	3.521E-10	3.189E-07	6.422E-10

由于 CGPA、GRE.Score 两者与普通残差的等级相关系数比较接近，经试验，我们选定 CGPA 构造权函数。并根据对数似然函数达到极大(AIC 达到最小)的原则选出最优的拟合指数 m=-10，于是构建加权最小二乘模型(fit.w1)如表 10 所示：

表 10 加权最小二乘模型(fit.w1)

	Estimate	Std.Error	t-value	Pr(>|t|)	
(Intercept)	-1.096	0.132	-8.328	0.000	***
GRE.Score	0.002	0.001	2.615	0.009	**
TOEFL.Score	0.003	0.001	2.516	0.012	*
University.Rating2	-0.014	0.022	-0.636	0.525	
University.Rating3	-0.012	0.023	-0.509	0.611	
University.Rating4	-0.001	0.025	-0.036	0.971	
University.Rating5	0.020	0.025	0.791	0.430	
SOP12	0.020	0.015	1.385	0.167	
SOP13	0.022	0.017	1.313	0.190	
LOR12	0.016	0.017	0.959	0.338	
LOR13	0.032	0.018	1.783	0.076	.
CGPA	0.113	0.012	9.058	0.000	***
Research1	0.039	0.008	4.609	0.000	***
RSE	2725		F: p-value	0.000	
R^2	0.841		Adjust R^2	0.835	

注意到幂指数 m 的取值与常规的取值范围[-2,2]偏差过大，模型 RSE 过大且与原始全模型相比较拟合优度的提升程度有限，于是尝试重新寻找权函数：先将原始全模型残差的绝对值作为因变量与所有自变量建立普通最小二乘回归模型，再将该模型拟合值的平方的倒数作为加权最小二乘的权重，建立新的加权最小二

乘模型(fit.w2)，如表 11 所示：

表 11 改进的加权最小二乘模型（fit.w2）

	Estimate	Std.Error	t-value	Pr(>|t|)	
(Intercept)	-1.001	0.103	-9.740	0.000	***
GRE.Score	0.001	0.000	3.035	0.003	**
TOEFL.Score	0.003	0.001	4.425	0.000	***
University.Rating2	-0.010	0.013	-0.791	0.429	
University.Rating3	-0.002	0.014	-0.161	0.873	
University.Rating4	-0.016	0.016	1.007	0.315	
University.Rating5	0.032	0.017	1.896	0.059	.
SOP12	0.027	0.013	2.002	0.046	*
SOP13	0.030	0.016	1.895	0.059	.
LOR12	0.035	0.013	2.668	0.008	**
LOR13	0.049	0.015	3.280	0.001	**
CGPA	0.104	0.009	12.007	0.000	***
Research1	0.028	0.008	3.381	0.000	***
RSE	1.371		F: p-value	0.000	
R^2	0.921		Adjust R^2	0.918	

该模型表明，在 0.05 的显著性水平下，除大学排名及目的陈述文案是否优秀外，其余变量均对申请者的录取几率有显著影响。

将加权最小二乘模型(fit.w1)、改进的加权最小二乘模型(fit.w2)与原始全模型比较，可见判决系数 R-square 增大，且图 10 所示的标准化残差图及表 12 的 ncvTest 都表明异方差问题被有效解决。

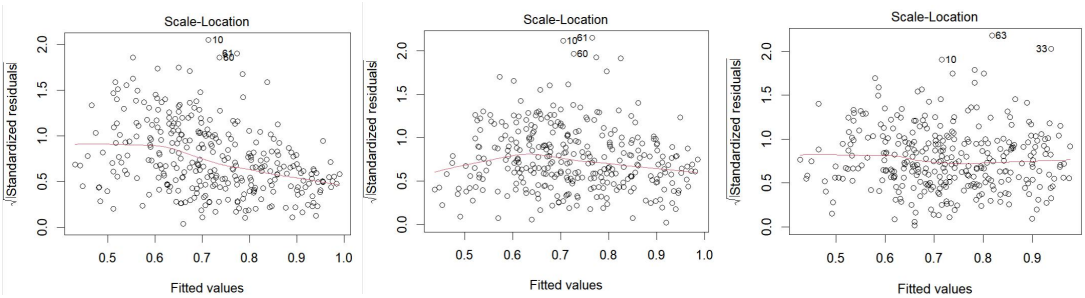

图 10 标准化残差图(从左至右：原始全模型、fit.w1、fit.w2)

表 12 ncvTest

	Chi-square	p-value
原始全模型	35.806	2.180E-09
fit.w1	1.809	0.179
fit.w2	0.087	0.768

如表 13 所示，AIC 值也表明模型获得提升：

表 13 AIC 值

	AIC
原始全模型	-841.544
fit.w1	-889.980
fit.w2	-944.025

故后续选择改进的加权最小二乘模型(fit.w2)做预测。

5.5 分位数回归

分位数回归又称 QR 回归，其原理是将数据按照因变量拆分成多个分位数点，研究不同分位数点情况下的回归影响关系情况。在分位数回归中，估计和推断是无分布的。分位数回归是线性回归的扩展，即当不满足同方差和正态性假定，也可以使用它。分位数回归得到一簇拟合曲线，对应不同的分位数水平。

分别在 0.05，0.25，0.50，0.75，0.95 五个分位数水平下做分位数回归，研究不同分位数水平下自变量与因变量的关系，系数估计如表 14 所示：

表 14 分位数回归结果

分位数	0.05	0.25	0.5	0.75	0.95
(Intercept)	-0.93	-1.67974	-1.03605	-0.81306	-0.512
GRE.Score	-0.0006	0.00363	0.00165	0.00066	0.00032
TOEFL.Score	0.0044	0.00191	0.00293	0.00408	0.0032
University.Rating2	-0.103	-0.04002	0.00517	0.02261	0.051
University.Rating3	-0.115	-0.02247	0.01588	0.02657	0.0444
University.Rating4	-0.0938	-0.03147	0.02594	0.04227	0.0465
University.Rating5	-0.0248	-0.00428	0.04712	0.05448	0.0545
SOP12	0.0049	0.03882	0.00822	0.02125	0.00051
SOP13	-0.02	0.03677	0.00477	0.01973	0.00756
LOR12	0.06	0.02107	0.03142	0.02385	-0.00994
LOR13	0.149	0.04038	0.04575	0.03314	0.00613
CGPA	0.142	0.11378	0.09996	0.09791	0.0933
Research1	0.0607	0.01674	0.0324	0.02451	0.0332

基于上述系数估计结果，可以比较不同分位数水平下各自变量与因变量的关系的差异。从中可以发现，不同分位数估计下的 CGPA 的系数，随着分位数的增加而降低。相较于录取率高的学生群体，录取率较低的学生群体的 CGPA、LOR 的对于录取率的影响更为明显，以本科院校评分 1 为基准，本科院校评分为 2、3、4、5 反而会不利于录取。

6.预测结果分析

除分位数回归模型以外，我们在本节比较其余多个模型间的预测结果。由于录取几率不存在 0 值，可以使用平均绝对百分比误差（MAPE）来衡量预测模型的准确程度。通过下述公式计算 MAPE，其中A_i为真实值，F_i为预测值：

$$MAPE = \left(\frac{100\%}{n}\right)\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right|$$

分别使用上述模型对样本量为 80 的预测集进行预测，结果如表 15 所示：

表 15 预测效果汇总

	全模型	AIC 选模型	BIC 选模型	LASSO 回归	回归树	加权最小二乘回归
MAPE	8.05%	8.04%	8.11%	8.29%	8.27%	7.75%

分析上表可以发现，预测效果最好的是加权最小二乘回归模型，其次是经过 Box-Cox 变换后的 AIC 选模型。预测效果相对较差的是 LASSO 回归和回归树。但是从整体上看，所有模型的预测效果都比较理想，前文中得出的结论均是可靠的。

7.结论与建议

7.1 留学竞争力提升策略

经过前文的各类分析，我们基于留学申请者的多个维度，分别给出针对性的意见，以此来提高他们在留学申请中的竞争力：

- 绝大部分模型结果表明，GRE 是影响留学竞争力的重要指标，在 GRE 上取得一个较高的分数，将会很大程度上提升留学竞争力。训练集的申请者 GRE 水平整体较高，中位数与平均值在 317 分左右。但一般公认的 GRE 优秀分水岭为 320 分，在现实情况中需要取得更高的分数，例如 330 分以上。
- 同时，TOEFL 成绩也是另一个值得重视的指标，部分模型认为其影响较 GRE 稍弱。分析原因可能是训练集申请者 TOEFL 成绩均值与中位数

都达到了 107 分。虽然托福成绩越高越好，但是考虑到其考试难度与考试成本，并且多数美国 TOP50 大学的托福基本要求为 100 分，我们建议 TOEFL 成绩略高于 100 分即可。

- 对于本科院校的档次而言，部分模型认为，录取几率只与申请者是否是最强一档院校有关，这符合我们的常识，因为部分国外知名院校有目标院校清单。根据分位数回归模型，如果某申请者的本科院校十分强劲，但是该申请者的软硬实力都较差，其录取几率会大打折扣，因此一个好的本科院校的确可以让申请者有一定的光环，但也要求申请者脚踏实地地努力学习，不断提高自己。
- 绝大部分模型认为，目的陈述并不重要，只有少部分模型认为，目的陈述是否为最高档才对录取几率有正向影响。所以，一份能够反映申请者强烈意愿以及真切情感的目的陈述文案才是好的文案。
- 对于 CGPA 而言，所有模型都认为这是极其重要的指标，因此申请者忙于参加各种活动，参加各种标化考试的时候，一定要保住自己的 CGPA。此外，根据分位数回归模型，我们可以发现，觉得自己的简历没有他人那么亮眼的申请者，完全可以花功夫提升自己的 CGPA，这是提升自己录取几率的性价比最高的方式。
- 推荐信档次越高越好，根据分位数回归模型，尤其是对于简历较为贫瘠，经历不那么丰富的申请者，一个更强的推荐信会带来更高的录取几率。
- 几乎所有模型都认为，有科研经历的申请者会比没有科研经历的申请者有更强的竞争力，因此，许多大学在本科阶段鼓励学生进行科研探索是有科学依据的。

7.2 回归分析方法总结与思考

与此同时，我们还对回归分析方法进行了总结与思考。

在进行回归分析的时候，数据的面貌往往不会符合理论假设，会产生异方差、非正态、复共线性等问题。我们需要有发现数据“病症”的能力，再根据“临床经验”对症下药。

建立回归模型不仅是一门科学，而且是一门艺术。目前没有模型能够完美地

解决所有的问题。提出一个新的模型，至少要确保模型在以下三个方面之一做得好：

1.参数估计精度，即模型的解释性方面。

2.预测精度，即预测集上的预测准确程度。

3.统计性质的符合程度，包括同方差、满足正态假定、独立性等。

Box-Cox 变换让数据重新满足统计假设；AIC、BIC 逐步回归能够得出拟合程度较高同时更加简洁的模型；LASSO 回归能够解决复共线性带来的参数估计不准问题；分位数回归通过非参数假设解决传统回归统计性质违背问题，同时给出更具有解释性的回归系数；加权最小二乘估计能够解决普通最小二乘模型的异方差性，并提升模型的拟合效果与预测效果；机器学习回归树模型有较好的预测效果，更简洁直观的判断方法，以及适用于多种类型的数据。

同样地，它们也有相应的缺点，例如由于基于不同的分位数给出回归方程，分位数回归对于数据整体解释能力不强；加权最小二乘回归在方差改进的方面可能对于某些数据不适用，普适性差于 box-cox 变换，对于权函数的确定可能会存在困难。总而言之，综合运用多种模型，模型间优势互补是有必要的。