Statistics and Application 统计学与应用, 2023, 12(4), 998-1008
Published Online August 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.124103

Hans汉斯

基于二元逻辑回归的考研动机影响因素分析

赵雪娇

北方工业大学理学院，北京

收稿日期：2023年7月17日；录用日期：2023年8月7日；发布日期：2023年8月22日

摘　要

近年来，关于“考研热”与“考研难”的话题频上热搜。为了帮助有考研动机的学生认清考研前景、做出理性选择，本文运用问卷调查广泛调研收集数据，采用核主成分分析对数据进行降维，构建二元逻辑回归模型深入挖掘考研动机是否明确的影响因素。研究结果表明，性别、疫情因素、提升就业竞争力三个因素对考研动机的影响最为显著。最后，本文从学生自身、家庭层面、院校人才培养模式、相关部门等多个层面提出对策建议。

关键词

考研动机，核主成分分析，二元逻辑回归，影响因素

Analysis of Influencing Factors of Postgraduate Entrance Examination Motivation Based on Binary Logistic Regression

Xuejiao Zhao

School of Science, North China University of Technology, Beijing

Received: Jul. 17th, 2023; accepted: Aug. 7th, 2023; published: Aug. 22nd, 2023

Abstract

In recent years, the topic of “entrance examination hot” and “entrance examination difficult” has been frequently searched on the Internet. In order to help students with the motivation to get a clear understanding of the prospect of postgraduate entrance examination and make a rational

文章引用: 赵雪娇. 基于二元逻辑回归的考研动机影响因素分析[J]. 统计学与应用, 2023, 12(4): 998-1008.
DOI: 10.12677/sa.2023.124103

choice, this paper extensively collects data through questionnaire survey, uses kernel principal component analysis to reduce the dimension of the data, and constructs a binary logistic regression model to deeply explore the influencing factors of whether the motivation is clear or not. The results show that gender, epidemic factors and improving employment competitiveness have the most significant impact on the motivation for postgraduate entrance examination. Finally, this paper puts forward countermeasures and suggestions from the aspects of students themselves, family level, college talent training mode and relevant departments.

Keywords

Motivation for Postgraduate Entrance Examination, Kernel Principal Component Analysis, Binary Logistic Regression, Influencing Factor

1. 引言

盲目追求高学历而引发的考研热，与教育的本质也是背道而驰的。随着考研的人数越来越多，学术界对考研热这一现象的研究逐步深入。就国内来看，国内学者的研究主要呈现两个特点。一是研究具体化，学者聚焦一所高校进行分析。二是研究领域化，学者聚焦某类专业进行研究。就国外来看，国外学者对影响研究生报考的因素进行了分析，这些影响因素包括个人所学专业、学术能力、劳动力市场情况、就业前景等。

本文在现有研究成果的基础上，通过问卷调查法收集数据，综合运用核主成分分析和二元逻辑回归分析法建立模型，选择定量和定性相结合的研究方法，对考研热现状及考研动机影响因素进行分析，最后根据研究结论提出建议。

2. 研究方案设计与实施

2.1. 问卷设计

本文通过问卷调查法来收集资料，调查对象为全国各地具有考研意愿的在校大学生及已经上岸的研究生，主要以线上发布问卷的形式进行数据的收集。调查问卷主要包括以下几个方面的内容：

1) 大学生基本信息调查(性别、年级、所学专业、本科阶段年级排名等)；
2) 大学生所在本科院校调查(本科学校类型)；
3) 具有考研意愿群体的具体情况调查(是否跨考、备考时间、考研信息获取途径、择校倾向等)；
4) 没有考研意愿群体的具体情况调查(不选择考研原因、对考研热的认识等)；
5) 大学生考研动机的影响因素调查。

2.2. 问卷实施

本次调查共计发放问卷 643 份，收回有效问卷 623 份，有效率为 97.36%。在调查时，为了提高问卷回收率，在问卷开篇向被调查者说明问卷资料的保密性和使用目的，问题的设计涉及多角度且又简洁易懂。正式调查之前先进行预调查，回收 150 份问卷后检验信效度，检验得信效度较高，可以进入正式调

查。然后在问卷星平台导出所收集到的数据，检验得：登录 IP 地址有 16 条记录重复，4 条记录用户作答时间过长。剔除无效填写记录后保证了问卷的有效性，因此可以利用此数据进行后续的问题分析。

2.3. 研究安排

1) 确定研究题目，查阅并梳理参考文献，明确研究内容及所使用的具体方法；编制资料搜集、整理以及分析研究计划。

2) 制作问卷，完成数据搜集工作，分析大学生考研意愿现状，完成描述统计分析。

3) 利用 Python、SPSS 等软件，完成数据分析及建模工作。

4) 整理分析结果，撰写论文的结论和建议，完成论文初稿。

5) 在指导老师指导下，对论文进行修改和完善。

6) 完成论文终稿，确认无误后打印装订。

3. 考研现状调查

3.1. 受访人群基本信息

根据被调查者的性别、所在年级、本科学校类型、所学专业等分别对样本进行大致分类汇总，样本分布情况见表 1。

Table 1. Basic information of interviewees
表 1. 受访人群基本信息表

类别	人数	百分比
男	272	42.30%
女	371	57.70%
大一、大二年级	224	34.84%
大三年级	174	27.06%
大四年级	152	23.64%
已经上岸	85	13.22%
其他	8	1.24%
有	536	87.58%
无	32	5.23%
模糊	44	7.19%
工学类	75	12.27%
理学类	114	18.66%
医学类	63	10.31%
法学类	73	11.95%
文学类	96	15.71%

Continued

艺术、教育、历史类	51	8.35%
经济、管理类	109	17.84%
其他	30	4.91%
不跨考其他专业或院校	267	43.70%
跨考其他专业或院校	225	36.82%
不清楚	119	19.48%

3.2. 考研现状分析

从考研动机上看，调查受访人群中 536 人有考研意愿，占总人数的 87.58%；44 人对是否考研不太确定，占总人数的 7.19%；32 人无考研意愿，仅占总人数的 5.23%。如图 1 所示。

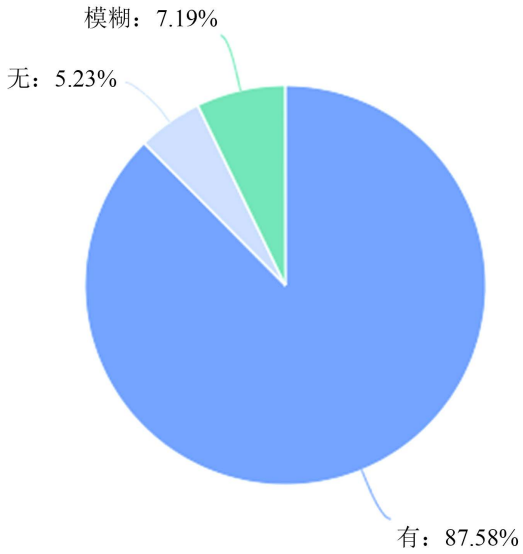

Figure 1. Intention of postgraduate entrance examination
图 1. 考研意愿

为了解主修专业与考研动机的关联，笔者运用 Excel 统计分析功能做主修专业和考研意愿交叉分析表，如表 2 所示。可得：医学类和艺术、教育、历史类两大类最具考研倾向，而主修工学类的学生考研动机最为轻微。现实生活中，医学生想要进入三甲医院工作，往往要求硕士、博士学历。除此之外对医学生毕业的院校也有要求，知名高校的毕业生更具优势。而艺术类在大学中可能尚未精确到某一发展方向，考研能在感兴趣的领域里获得更多的知识储备。因此，考研成为医学类和艺术、教育、历史类两大类专业的最佳选择。

此外，在择校优先度上，对受访群体进行调查并设计排序题，如图 2 所示。排序题的选项平均综合得分是由问卷星系统根据所有填写者对选项的排序情况自动计算得出的，得分越高表示综合排序越靠前。具有考研动机的学生在择校时最为看重对今后发展有利的院校以及所报考专业实力较强的学校优先。归根结底，还是为了提升自身就业竞争力。

Table 2. Cross-analysis table of undergraduate major and postgraduate entrance examination intention
表 2. 本科主修专业和考研意愿交叉分析表

X\Y	有	无	模糊	小计
工学类	62 (86.11%)	0 (0.00%)	10 (13.89%)	72
理学类	102 (91.89%)	0 (0.00%)	9 (8.11%)	111
医学类	62 (98.41%)	0 (0.00%)	1 (1.59%)	63
法学类	65 (92.86%)	0 (0.00%)	5 (7.14%)	70
文学类	84 (91.30%)	0 (0.00%)	8 (8.70%)	92
艺术、教育、历史类	45 (95.74%)	0 (0.00%)	2 (4.26%)	47
经济、管理类	86 (90.53%)	0 (0.00%)	9 (9.47%)	95
其他	30 (100%)	0 (0.00%)	0 (0.00%)	30

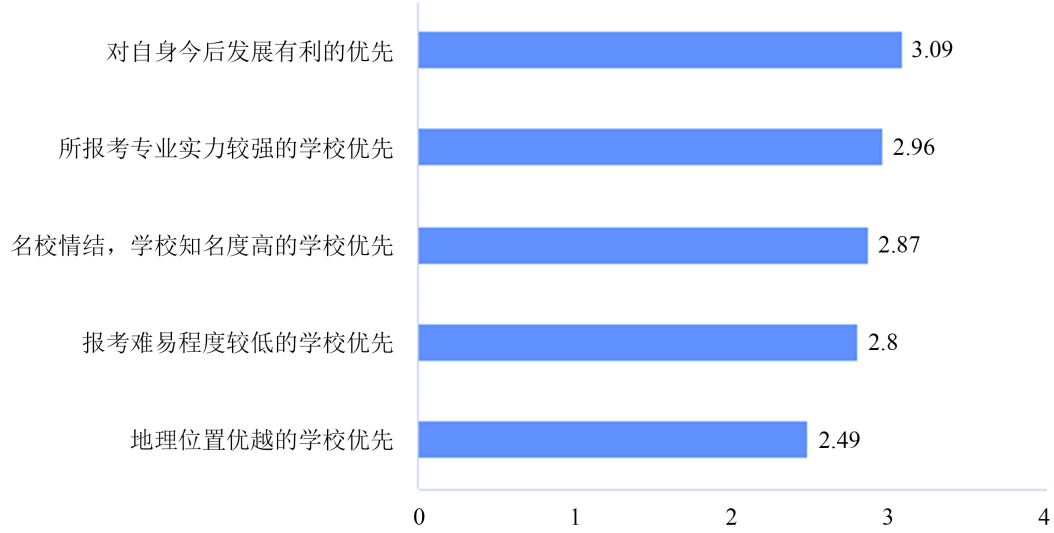

Figure 2. Ranking of school selection priorities
图 2. 择校优先程度排序

4. 探索性数据分析

4.1. 探索性数据分析相关概念介绍

探索性数据分析(Exploratory Data Analysis, EDA) [1]是数据深入挖掘的前提准备和必要参与。一方面它可以查看数据是否存在异常值、缺失值，为数据预处理打基础；另一方面，它可以对数据的整个分布和特征之间的关系进行预览，预览结果直接通过各种图表(例如：条形图、箱线图、热力图等)将结果可视化，直观明了，从而找到变量间的深层关系，获取模型构建的思路。

4.2. 特征间的相关关系

为观察数据集各个特征之间的关系，并观察各个特征与考研动机之间的相关性强弱。我们借助 Python 中极为常用的热力图来实现这个问题。热力图通过颜色的深浅来表示数据的分布。如图 3 所示。

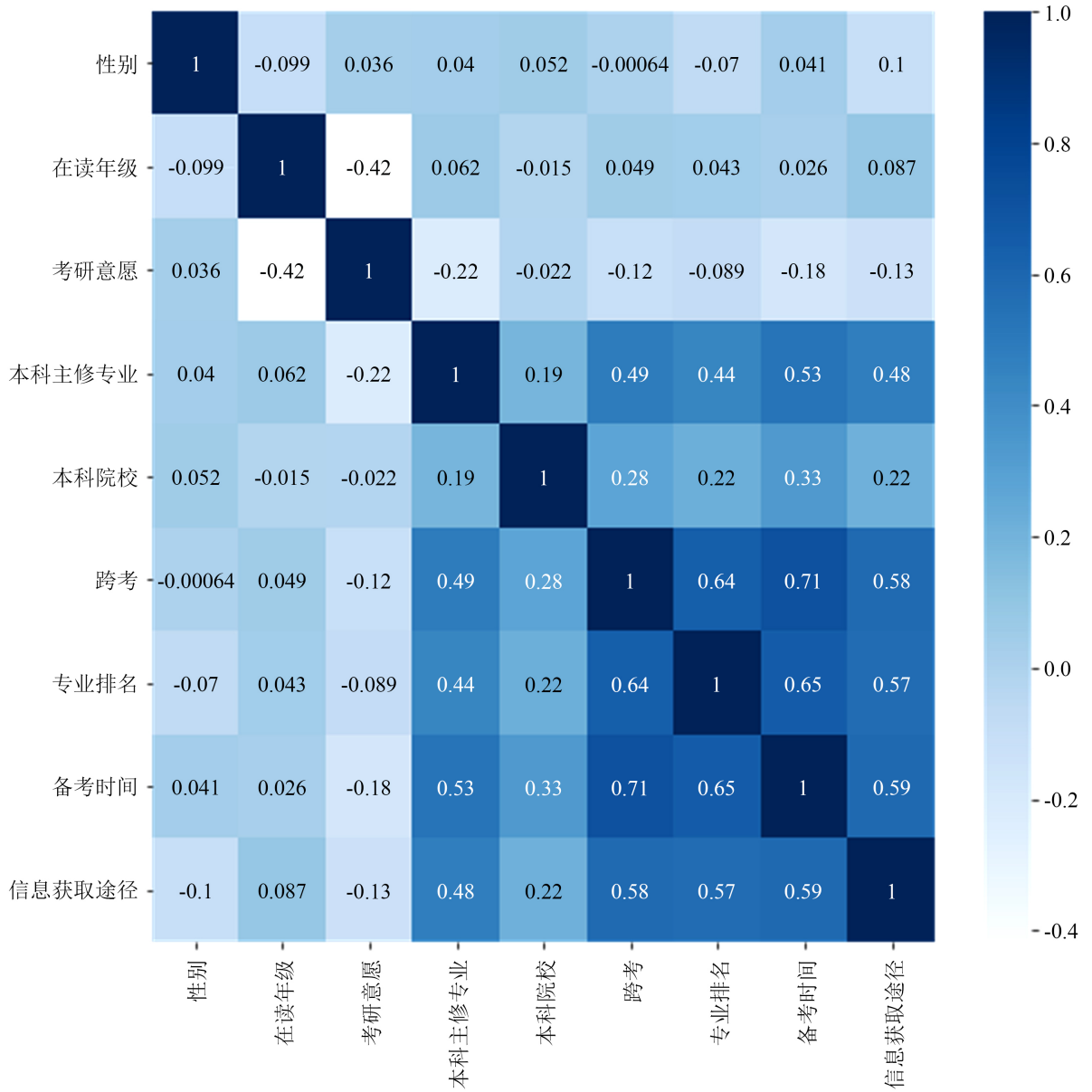

Figure 3. Heat map of characteristic analysis
图 3. 特征分析热力图

5. 对考研动机的影响因素研究

5.1. 降维算法

主成分分析(PCA)适用于数据的简单线性降维。但是当遇到难以使用线性变换降维的数据时，采用PCA 降维会丢失原本的低维结构。而核主成分分析(KPCA)，不会丢失原本的结构，能够较好地进行数据的非线性降维，适用于线性不可分的数据集的处理。

KPCA 的基本思想：先将原始空间中待降维的数据，采用映射函数变换到高维甚至是无穷维的空间，再把数据从高维空间进行 PCA 降维，投影降维到需要的维数。数据被映射至高维，原本线性不可分的数据集就会变的线性可分了。核主成分分析法通过核函数来完成计算，原理如图 4 所示。

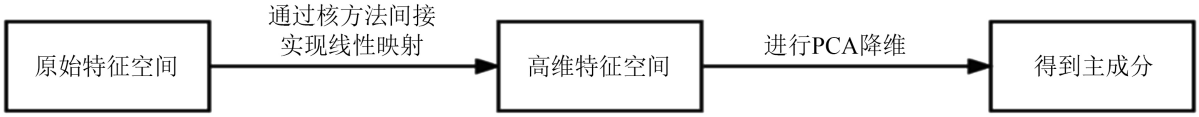

Figure 4. Schematic diagram of KPCA
图 4. KPCA 原理图

5.2. KPCA 算法步骤及降维结果

算法步骤如下：

① 通过核函数计算矩 $K=\left\{k_{ij}\right\}_{n\times n}$，其元素为 $K_{ij}=k\left(x_i,x_j\right)$。其中 x_i 和 x_j 为原空间的样本，$k(\cdot,\cdot)$ 是核函数。

② 计算 K 的特征值，并按照由大到小的顺序排列。找出由特征值对应的特征向量 α^l (表示第 l 个特征向量)，并对 α^l 进行归一化($\left\|\alpha^l\right\|=1$)。

③ 原始样本在第 l 个非主成分下的坐标为：$z^l(x)=\sum_{i=j}^{n}\alpha_i^l k\left(x_i,x\right)$

这里的 x_i 是指第 i 个样本，α^l 的维度与样本数相同。如果计算 K 的前 m 大个特征值及相应的特征向量，则样本在前 m 个非线性主成分上的坐标就构成了在新空间中的$\left[Z^1(x),Z^2(x),\cdots,Z^m(x)\right]^{\mathrm{T}}$表示。

④ 通过对比分析各特征的重要性比例，剔除关联性较低的特征，并选取相关特征进行降维操作，由此提取对影响考研动机较有价值的特征，提高后续模型的训练效率并提升模型的准确度。

降维前，影响考研动机的有 11 个维度，分别为：性别、本科主修专业、在读年级、逃避就业、不满意本专业、提升就业竞争力、获得社会认同感、渴望投身科研事业、父母的期望、考研大趋势和疫情因素不便出国留学。

KPCA 降维后，影响考研动机的主要因素有 6 个维度，分别为：性别、提升就业竞争力、获得社会认同感、渴望投身科研事业、疫情因素不便出国留学和考研大趋势。

5.3. 逻辑回归模型

5.3.1. 模型介绍

逻辑回归[2] [3]是一个假设样本服从伯努利分布，利用极大似然估计和梯度下降求解的二分类模型，在分类、CTR 预估领域有着广泛的应用。逻辑回归与线性回归不同的是，逻辑回归输出的不是具体的值，而是一个概率。逻辑回归不能用于连续数值的预测，它的好处在于逻辑回归的模型建立起来之后，对于大量新数据的分类的计算都会变得十分快速且消耗的计算资源很少。逻辑回归的决策边缘是线性的，因此逻辑回归是线性模型。

5.3.2. 模型构建

逻辑回归模型构建的基本思路[4]：1) 首先获取各个类的 propensities (属于该类的概率)，比如 Y = 1 的 propensity 就是 $p=P(Y=1)$；2) 然后设置 cutoff value，并根据该值同 propensities 比较的结果，将新数据分类，比如若 $p=P(Y=1)<0.5$，则分类为 0，否则为 1。

逻辑回归的设计思路：从逻辑回归的设计目的开始，逻辑回归的设计主要是为了分类，multilinear regression model 输出的是一个连续的预测值 Y (其范围可能是(−inf, +inf))，而预测值 Y 同分类问题结合起来就要借助 logit 函数。根据这个设计思路，逻辑回归模型的公式如下：

$$\log it=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_dx_d \tag{1}$$

$$p=\frac{1}{1+\mathrm{e}^{-\log it}}=\frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_dx_d)}} \tag{2}$$

以考研动机是否明确 Y 为因变量，性别(x_1)、提升就业竞争力(x_2)、疫情因素不便出国留学(x_3)、考研大趋势(x_4)、渴望投身科研事业(x_5)和获得社会认同感(x_6)六个维度为自变量，通过 SPSS 进行二元 logistic 回归模型分析拟合。

5.3.3. 模型结果与分析

以考研动机是否明确 Y 为因变量，六个维度为自变量构建二元 logistic 回归模型，结果如表 3~7 所示。

Table 3. Model summary
表 3. 模型汇总

模型摘要				
−2 对数似然	考克斯－斯奈尔 R 方	内戈尔科 R 方	卡方	显著性
247.825[a]	0.290	0.224	55.959	0.000

由表 3 可知，卡方值为 55.959，P 值为 0.000，小于 0.01，通过了显著水平为 1%的显著性检验，由此可知模型具有统计学意义。此外，−2 对数似然值为 247.825，考克斯－斯奈尔 R 方为 0.290，内戈尔科 R 方为 0.224，由此可知，模型的拟合度较高，即模型对原始数据的解释度比较理想。

Table 4. Hosmer-Lemeshaw test
表 4. 霍斯默－莱梅肖检验

卡方	自由度	显著性
11.939	8	0.154

由表 4 霍斯默－莱梅肖检验结果可知，卡方值为 11.939，P 值为 0.154，大于 0.05，说明模型的预测的拟合度良好。

Table 5. Variables in the equation
表 5. 方程中的变量

	B	标准误差	瓦尔德	自由度	显著性	Exp(B)	EXP(B)的 95%置信区间	
							下限	上限
性别 x_1	−1.176	0.475	6.14	1	0.043	0.308	0.122	0.782
提升我的就业竞争力 x_2	−0.454	0.263	2.968	1	0.013	0.635	0.379	1.064
疫情等不确定因素不便出国留学 x_3	0.137	0.164	0.699	1	0.041	1.147	0.832	1.581
考研的大趋势 x_4	0.221	0.201	1.206	1	0.272	1.247	0.841	1.85
热爱研究，渴望投身科研事业 x_5	0.873	0.206	18.032	1	0.394	2.394	1.6	3.582
得到社会认同感 x_6	0.257	0.233	1.212	1	0.271	1.293	0.818	2.042
常量	1.147	1.276	0.808	1	0.369	3.149		

由表 5 可知，性别 x_1、提升就业竞争力 x_2、疫情因素不便出国留学 x_3 三个指标对考研动机是否明确 Y 存在显著的影响($p < 0.05$)。其中，性别 x_1 的系数为−1.176，瓦尔德卡方值为 6.14，p 值为 0.043，小于 0.05，通过了显著水平为 5%的显著性检验；提升就业竞争力 x_2 的系数为−0.454，瓦尔德卡方值为 2.968，p 值为 0.013，小于 0.05，通过了显著水平为 5%的显著性检验；疫情因素不便出国留学 x_3 的系数为 0.137，瓦尔德卡方值为 0.699，p 值为 0.041，小于 0.05，通过了显著水平为 5%的显著性检验。

根据系数表构建模型如下：

$$\log it(p) = 1.147 - 1.176x_1 - 0.454x_2 + 0.137x_3 \tag{3}$$

Table 6. Forecast table
表 6. 预测表

实测			预测		
			考研意向明确与否		正确百分比
			0	1	
步骤 1	考研意向明确与否	0	2	40	4.8
		1	3	551	99.5
	总体百分比				92.8

以 0.5 的概率为分界线，采用上述模型进行预测，结果如表 6 所示。其中预测为考研意向不明确，现实考研意向不明确的样本有 2 位，预测正确率为 4.8%；预测为考研意向明确，现实考研意向明确的样本有 551 位，预测正确率为 99.5%。由此可知模型在预测具有考研意向时预测结果非常理想。以预测概率与真实值构建 ROC 曲线如图 5 所示。

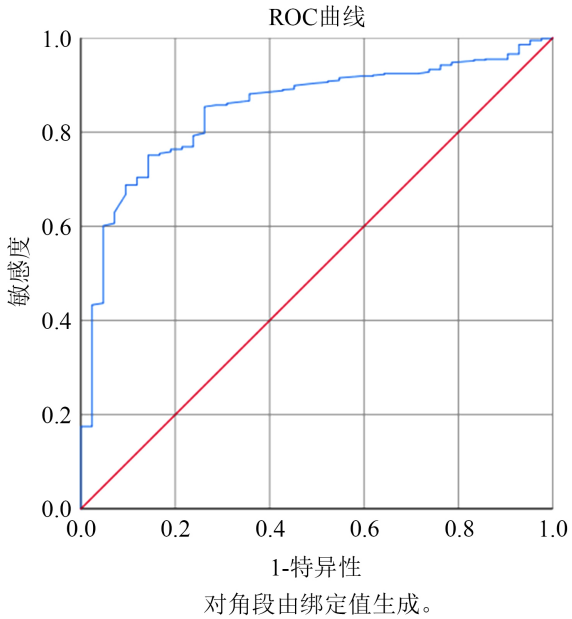

Figure 5. ROC curve
图 5. ROC 曲线

由表 7 知，其下方所包面积为 0.849，说明该模型对于目标变量具有很好的预测效果。根据其敏感度和特异性计算得到的尤登指数最高为 0.608，其最佳分界点为 0.932。

Table 7. The area below the curve
表 7. 曲线下方的区域

区域	标准错误 [a]	渐近显著性 [b]	渐近 95%置信区间	
			下限	上限
0.849	0.026	0.000	0.798	0.900

6. 结果讨论

6.1. 研究结论

本文采用核主成分分析方法对数据原始指标进行降维处理后，提取到 6 个对影响考研动机较有价值的特征，分别为：性别、提升就业竞争力、获得社会认同感、渴望投身科研事业、疫情因素不便出国留学和考研大趋势。再通过 SPSS 进行逻辑回归拟合，结果表明疫情因素不便出国留学、提升就业竞争力、性别三个因素对考研动机的影响最为显著。

本文采用定量和定性相结合的方法对考研动机主要影响因素进行研究，具有较高的科学性和可参考性，有助于帮助具有考研动机的学生认清考研前景、做出理性选择，引导学生树立正确的人才观、就业观。此外，该方法还可以应用于研究毕业生就业能力，学生入党，素质测评以及心理健康等方面[5]。

6.2. 建议

首先，学生本人应尊重自己的内心，并做好职业规划，而不是盲目跟风考研。高学历并不等于高能力，不能只贪图轻松、稳定的工作，要树立诚实劳动、踏实肯干的劳动观，“人尽其才，才尽其用”的人才观。家庭层面，父母应充分尊重子女选择，可以提供建议但不强加干预。政府相关部门也应制定、完善相关法律法规，加强监管，消除“学历歧视”，完善大学毕业生就业政策，确定合理的研招规模。此外，院校应提升大学生的培养质量，正确引导考研风气，帮助大学生走出学历与文凭的误区，让笼罩在考研上的迷雾消散，才能有效缓解当下的盲目的考研热情。

7. 结语

综上所述，“考研热”表明就业形势的变化和社会高学历人才需求上涨的同时，也应引起社会层面的深入思考。当务之急恐怕不是给这个数字本身降温，而是做“大蛋糕”，让考研变得不再刚需。上岸不是尽头，未来不会一直搁浅。学校是一座象牙塔，就业市场更是没有硝烟的战场，无论是为了减轻就业压力，亦或是提升学历，甚至是盲从者，其实人生的每个路口都会有风景等待欣赏，考上研究生的同学不要自我松懈，没考上也不用过度自责，研究生考试并非人生的全部。

致　谢

本篇论文我有幸受到徐老师的指导进行考研动机影响因素研究，我在项目研究过程中收获良多。从选题到确定主要模型，再到成稿，我们不断在矛盾中斟酌，在疑惑中探索。

首先感谢指导教师徐老师，她在我们的选题、收发问卷、研究思路方面提供了非常宝贵的建议和指导。还要感谢我的同学、朋友及家人，积极地帮我填写、转发问卷。最后，对参加本论文审阅和对本论

文提出宝贵意见的老师表示诚挚的谢意。

参考文献

[1] David C. Hoaglin, Frederick Mosteller, John W. Tukey. 探索性数据库分析[M]. 陈忠琏, 郭德媛, 译. 北京: 中国统计出版社, 1998, 11-32.

[2] 邹媛. 二元逻辑回归模型中几类一阶近似刀切估计的研究[D]: [硕士学位论文]. 贵阳: 贵州民族大学, 2021.

[3] 陈敏. 逻辑回归模型中样本量确定的相关问题研究[D]: [硕士学位论文]. 昆明: 云南大学, 2020.

[4] 杜谦, 范文, 李凯, 杨德宏, 吕佼佼. 二元Logistic回归和信息量模型在地质灾害分区中的应用[J]. 灾害学, 2017, 32(2): 220-226.

[5] 李斌. 高校学生综合素质测评的模糊综合评价[J]. 商洛学院学报, 2017, 31(2): 76-79.