doi:10.3969/j.issn.1001-4616.2026.03.014

基于可解释机器学习的科技人才流动宏观影响因素研究

徐倪妮,孟　涵,朱淑婉

(合肥工业大学管理学院,安徽 合肥 230009)

[摘要]　为精准识别影响科技人才流动的重要因素及其作用方式,利用 1978—2023 年中国 31 个省(直辖市、自治区)的面板数据,构建基于 XGBoost 的科技人才流动预测模型,并引入 SHAP 可解释性方法完成特征影响全局解释、特征依赖分析、特征交互分析及分区域解释. 结果表明,科技人才流动受科研环境影响最为显著,其次为教育环境、经济发展水平、收入水平和生活便利度,产业结构的影响则相对较弱;科研环境与科技人才流动呈现正向线性关联,而其余因素则表现出复杂非线性关系;经济发展水平与科研环境、科研环境与生活便利度之间存在显著协同效应;东部、中部、西部及东北地区的科技人才流动的主导因素存在显著区域差异. 据此,提出促进区域科技人才合理流动的政策建议.

[关键词]　可解释机器学习,科技人才,人才流动,影响因素
[中图分类号]C962;TP181　[文献标志码]A　[文章编号]1001-4616(2026)03-0127-10

Research on Macro-Level Factors Influencing Science and Technology Talent Mobility Using Explainable Machine Learning

Xu Nini,Meng Han,Zhu Shuwan

(School of Management,Hefei University of Technology,Hefei 230009,China)

Abstract:To accurately identify key influencing factors and their impact mechanisms on science and technology talent mobility,this study employed panel data of China's 31 provinces(municipality,autonomous region)from 1978 to 2023 to construct an XGBoost-based prediction model for science and technology talent mobility. The SHAP interpretation method was then introduced to conduct global interpretation of feature importance,feature dependency analysis,feature interaction analysis,and regional interpretation. The results indicate that:science and technology talent mobility is most significantly influenced by the research environment,followed by the education environment,economic development level,income level,and living convenience,while the impact of industrial structure is relatively weak;The research environment exhibits a positive linear relationship with science and technology talent mobility,whereas other factors demonstrate complex nonlinear relationships;Significant synergistic effects exist between economic development level and research environment,as well as between research environment and living convenience;Significant regional variations exist in the dominant drivers of talent mobility across eastern,central,western,and northeastern China. Based on these findings,policy recommendations are proposed to promote the rational mobility of science and technology talents among regions.

Key words:explainable machine learning,science and technology talents,talent mobility,influence factor

在知识经济蓬勃发展和全球科技博弈升级的背景下,科技人才作为科技创新的核心要素,已成为不同国家和地区竞相争夺的稀缺资源. 党的二十大报告明确指出"必须坚持科技是第一生产力、人才是第一资源、创新是第一动力",并强调"加快建设世界重要人才中心和创新高地,促进人才区域合理布局和协调发展". 科技人才的合理有序流动是促进区域间知识互动、推动技术融合创新的重要途径,将深刻影响区域创新格局的演

收稿日期:2025-08-08.
基金项目:安徽省哲学社会科学规划青年资助项目(AHSKYQ2023D014).
通讯作者:徐倪妮,博士,讲师,研究方向:大数据管理与应用. E-mail:xunini@hfut.edu.cn

变与国家科技强国建设进程. 然而,当前我国科技人才流动仍面临区域分布失衡、跨领域流动活力不足等现实挑战,其形成与演化是宏观环境要素与个体选择行为动态交织的结果. 究竟哪些宏观因素是影响科技人才流动的核心驱动力? 这些因素对科技人才流动的影响是否存在复杂的非线性关系,而非简单的线性增减? 不同区域科技人才流动的驱动机制是否存在显著差异? 系统回答这些问题,不仅有助于丰富和发展人才流动理论,更能为优化区域人才政策及促进跨区域协同发展提供实证依据,对深入实施新时代人才强国战略具有重要的理论和现实意义.

已有文献围绕科技人才流动的影响因素已经进行了多维度研究. 从影响因素的层次划分来看,已有研究关注了宏观、中观、微观不同层面因素对科技人才流动的影响. 例如,在宏观层面,有研究证实区域经济发展水平、科技政策环境、生活便捷度、产业转型、开放包容的社会文化氛围是吸引科技人才的主要因素[1-4];中观组织层面,实证分析表明薪酬福利、工作环境、分配公平、职业发展与研发环境等会显著影响科技人才的流动意愿[4-6];微观个体层面,现有研究聚焦于年龄、性别、家庭生活和个人情感等特征如何影响流动决策[7-9]. 其次,一部分文献针对特定类型科技人才群体的流动动因展开了探讨. 针对高校科技人才,王冲[10]运用灰色关联分析,系统考察了宏观经济运行态势、产业结构特征、财政科技投入强度、劳动报酬竞争力及研究环境质量等因素与其流动性的关联度. 在企业科技人才领域,郭俊华等[5]通过层次回归分析,验证了分配公平、职业发展与研发环境对人才流动意向的负向抑制效应,而工作压力与程序化则会正向驱动流动. 对于科技智库人才,刘杨等[8]利用泊松回归和 logistic 回归方法,揭示了职业年龄、性别等变量与其流动频次间的统计关联. 此外,还有一部分文献重点研究了特定因素或几个因素组合对科技人才流动的影响及其差异. 例如,柳美君等[11]使用多期双重差分模型评估了长三角区域一体化政策的异质性效应,指出该政策显著促进了区域内城市间及城市内部的科技人才流动,其提升作用在自然科学领域、非相邻城市间及省际城市间的科技人才流动中尤为凸显. 王焘等[12]则从政策、科技和产业三维度解构区域新兴产业创新生态系统,并基于组态视角总结出多重创新生态要素组合驱动科技人才流动的典型路径,分别是全能型均衡驱动、科技要素主导型驱动以及政策引领与产业促进协同型驱动.

现有研究虽在科技人才流动影响因素的识别与分析上取得了重要进展,但仍存在一些亟待弥补的局限. 其一,多数研究仍停留在验证各因素是否产生显著影响,未能进一步揭示不同因素的相对重要性. 这一局限主要源于面板回归、层次回归等传统实证方法的技术特性,即在多因素对比分析中难以直接、一致地评估各变量的边际贡献,导致决策参考的针对性不足. 其二,对于因素间复杂的非线性关系与交互效应缺乏有效探测手段. 经济发展水平等因素对科技人才流动的影响很可能并非简单的线性关系,然而既往研究大多采用线性模型进行简化分析,即便少数研究尝试非线性建模,也需预先设定函数形式,可能因模型误设引发估计偏误,难以准确刻画真实的作用机制.

机器学习技术为突破上述局限提供了方法支撑. 其核心优势体现在三方面:第一,凭借优异的高维变量处理能力,可纳入更全面的影响因素体系;第二,数据驱动的特性使其无需预设函数形式,而是通过算法自主学习数据中的潜在模式,能够更精准地捕捉各因素与科技人才流动间的非线性关联及交互效应;第三,当与 SHAP(Shapley Additive Explanations)等可解释性方法结合时,能够实现对因素影响的严格量化解读. SHAP 方法基于博弈论中的合作博弈思想,可为每个特征在任一预测中的贡献提供一致且公平的数值估计,这使得研究能够突破传统仅关注系数显著性的局限,转向对变量进行全局重要性排序、分析特定变量的边际影响规律,探测多因素间的交互效应,并识别不同子群体中的异质性影响,从而在统一框架下实现“是否影响”、“何种程度影响”以及“如何影响”的综合解答.

值得注意的是,近年来机器学习已被逐步应用于人才领域的多个方向,包括人才吸引政策优化[13]、招聘效率提升[14]以及员工流失预测[15]等. 然而,系统性地将机器学习与可解释性技术结合,应用于宏观层面科技人才流动影响因素的辨识、重要性量化及复杂作用机制解析的研究尚显不足. 鉴于此,本文将基于 1978—2023 年中国 31 个省(直辖市、自治区)的面板数据,构建机器学习预测模型,并重点依托 SHAP 可解释性框架进行四个层次的深入分析:全局特征重要性评估、特定因素的依赖关系剖析、多因素交互作用的探测以及分区域的异质性解释,旨在识别影响科技人才流动的关键宏观因素,精准刻画其影响程度、作用方向及具体方式,厘清各因素效应的区域差异,从而为区域人才战略精准施策提供更具针对性的科学依据.

1 研究设计

1.1 研究方法

1.1.1 XGBoost 方法

XGBoost 是由 Chen 等[16]提出的一种高效梯度提升决策树(GBDT)改进框架,其核心思想在于通过加法模型与前向分步算法,串行地集成多棵决策树,其中每一棵新树的学习目标都是拟合前一棵树所遗留的预测残差. 这种基于决策树集成的模型结构,使其本质上具备强大的非线性拟合能力. 此外,XGBoost 通过引入二阶泰勒展开与并行计算策略,极大地提升了训练效率. 其出色的高维异质数据处理能力,能够有效捕捉宏观因素间复杂的非线性关系与高阶交互效应,这为后续深入探究多维度因素对科技人才流动的差异化影响提供了可靠基础. 在具体训练过程中,假设 $\boldsymbol{X}$ 是省际科技人才流动影响因素的宏观指标特征集合,$\boldsymbol{Y}$ 为各省科技人才流动量的观测值,给定一个包含 n 个样本的训练数据集 $D=\{(\boldsymbol{x}_i,y_i)\}(|D|=n,\boldsymbol{x}_i\in\mathbb{R}^m,y_i\in\mathbb{R})$,则集成模型的最终输出为 K 棵树的预测结果之和:

$$\hat{y}_i=\phi(\boldsymbol{x}_i)=\sum_{k=1}^{K}f_k(\boldsymbol{x}_i),\quad f_k\in\mathcal{F}$$

其中,$\hat{y}_i$ 是第 i 个样本的预测值,f_k 代表第 k 棵决策树,而 $\mathcal{F}$ 为所有可能的决策树组成的函数空间.

模型通过最小化以下正则化目标函数 $\mathcal{L}$ 来进行学习:

$$\mathcal{L}(\phi)=\sum_{i=1}^{n}\iota(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k).$$

其中,$\Omega(f)=\gamma T+\frac{1}{2}\lambda\boldsymbol{\omega}^2$.

该目标函数的第一部分 $\sum_{i=1}^{n}\iota(y_i,\hat{y}_i)$ 是衡量模型预测值与真实值之间差异的可微损失函数;第二部分 $\sum_{k=1}^{K}\Omega(f_k)$ 则是 XGBoost 的关键创新,用于控制模型复杂度的正则化项. 此正则化项通过惩罚复杂的树结构和过大的节点得分,有效提升了模型的泛化能力并显著降低了过拟合风险.

1.1.2 SHAP 可解释方法

虽然 XGBoost 在预测精度上表现优异,但其树结构的复杂性导致特征贡献难以直接量化. 因此,本文将进一步采用 SHAP(Shapley Additive Explanations)可解释性方法[17]. 该方法源于合作博弈论中的沙普利值(Shapley Value),其核心思想是将模型的最终预测结果视为所有特征参与者“协作”的产物,并通过公平分配来量化每个特征的边际贡献. 假设第 i 个样本的特征向量为 $\boldsymbol{x}_i$,其第 j 个特征为 x_{ij},XGBoost 模型对该样本的预测值为 $\hat{y}_i$,全体样本预测的基准值(即模型在所有特征缺失情况下的期望输出)为 ϕ_0. 根据 SHAP 的加性解释原理,模型的预测结果可分解为:

$$\hat{y}_i=\phi_0+\sum_{j=1}^{m}\phi_j(\boldsymbol{x}_i),$$

其中,$\phi_j(\boldsymbol{x}_i)$即为特征 j 在该样本预测中对应的 SHAP 值. 其数学本质是通过计算该特征在所有可能的特征子集中出现的边际贡献均值来求得:

$$\phi_j(\boldsymbol{x}_i)=\sum_{S\subseteq\{1,\cdots,m\}\setminus\{j\}}\frac{|S|!\ (m-|S|-1)!}{m!}[f(S\cup\{j\})-f(S)],$$

式中,S 是特征子集,$f(S)$是使用子集 S 中的特征进行预测的模型输出. 该值在数值上可正可负,直观反映了某一特征相较于基准预测对最终结果是起到了正向提升还是负向抑制作用.

相较于特征置换法(Feature Permutation)和局部可解释模型(LIME)等传统解释技术,SHAP 方法具备坚实的博弈论公理基础,其满足的一致性(Consistency)保证了对重要特征的评估更为可靠,并且对特征间的多重共线性也表现出较强的稳健性. 通过特征概要图(Summary Plot)、特征依赖图(Dependence Plot)及交互作用图(Interaction Plot)等可视化工具,SHAP 能够从全局与局部两个层面透明化“黑箱”模型的决策逻辑. 在本研究中,首先,通过全局 SHAP 值均值量化各因素的相对贡献度,明确其重要性排序,并借助特征概要图揭示

各因素与科技人才流动的总体相关性方向,区分促进或抑制作用;其次,通过特征依赖图解析各因素影响的线性、单调或复杂的非线性模式,刻画特征与目标变量间的动态关系;再次,通过分析特征间的交互效应,揭示多因素耦合对科技人才流动产生的协同或拮抗作用;最后,通过构建东部、中部、西部、东北四大经济区域独立的特征概要图,进行分区域解释分析,识别主导不同区域科技人才流动的差异化核心因素.

1.2　数据来源与变量测量

本文以 1978—2023 年中国 31 个省(自治区、直辖市)面板数据为研究样本,数据来源于《中国统计年鉴》、各省份统计年鉴及国泰安数据库,经多源交叉校验确保数据一致性.

针对科技人才流动的测度,借鉴既有研究[18-19],采用修正后的引力模型进行量化. 该模型通过整合工资水平、科研投入及地理距离等多维要素,克服了传统人口增长指标或单一人才存量统计的局限性. 具体测度公式如下:

$$TP_{ij}=R\&DP_i\times\frac{Wage_j}{Wage_i}\times\frac{R\&DE_j}{R\&DE_i}\times d_{ij}^{-b}, \tag{1}$$

式中,$R\&DP_i$ 代表省份 i 的科技人才存量,以规模以上工业企业的 $R\&D$ 研发人员数量表征;$Wage_j/Wage_i$ 为省份 j 与 i 的城镇单位就业人员年平均工资比值,用于刻画区域间收入势能差;$R\&DE_j/R\&DE_i$ 为省份 j 与 i 的规模以上工业企业 $R\&D$ 经费内部支出比值,体现创新资源投入的区域差异. d_{ij}为省份 i 与 j 之间的地理距离,b 为距离衰减指数,此处取 $b=2$,表明随着地理距离增加,科技人才流动的阻力会以平方级增长. 最终,省份 j 的科技人才吸引力通过汇总所有其他省份流向 j 的 TP_{ij}值,并加总得到总流入强度:

$$TP_j=\sum_i^n TP_{ij}. \tag{2}$$

基于人才流动理论与国内外相关研究[20-22],区域间经济、科技、教育、公共服务等维度的非均衡性是科技人才流动的重要驱动力. 本文据此构建包含 6 个核心维度的宏观影响因素指标体系(见表 1):(1)经济发展水平是区域综合实力的基础性指标,反映宏观经济环境对科技人才流动的支撑能力,采用人均 GDP 的自然对数值衡量;(2)收入水平直接关联人才的物质回报预期,是影响流动决策的关键因素,采用人均可支配收入的自然对数值衡量;(3)产业结构反映区域经济形态与产业升级方向,对科技人才需求类型具有导向作用,采用第三产业产值占 GDP 的比重衡量;(4)科研环境是科技人才职业发展的核心载体,直接影响其科研条件与创新机会,采用 R&D 经费支出占 GDP 的比重衡量;(5)教育环境通过人才培养与知识溢出影响科技人才流动的长期趋势,以教育支出占财政总支出的比重作为代理变量;(6)生活便利度是影响科技人才定居意愿的重要软环境,考虑到固定资产投资额反映了基础设施建设的规模与速度,以固定资产投资额的自然对数值作为代理变量.

表 1　宏观影响因素指标体系

Table 1　Macro influencing factors indicator system

因素名称	具体指标	参考来源
经济发展水平	人均 GDP 的自然对数值	
收入水平	人均可支配收入的自然对数值	窦旺胜等(2023)[20]
产业结构	第三产业产值占 GDP 的比重	徐倪妮等(2017)[21]
科研环境	R&D 经费支出占 GDP 的比重	王焘等(2024)[12]
教育环境	教育支出占财政总支出的比重	齐宏纲等(2025)[22]
生活便利度	固定资产投资额的自然对数值	

2　实证结果分析

2.1　模型构建与评估

在科技人才流动预测模型构建过程中,以经济发展水平、收入水平、产业结构、科研环境、教育环境、生活便利度等宏观因素作为特征变量,通过机器学习框架建立预测模型. 为保障模型泛化能力并避免过拟合,采用随机等比例抽样的方式将数据集划分为 80%的训练集与 20%的测试集. 分别采用 XGBoost、Random Forest、MLP、KNN、SVR 五种典型机器学习算法在训练集上训练科技人才流动预测模型,各模型均经过参数优化:

XGBoost 通过网格搜索确定最大树深度为 6、学习率为 0.1;随机森林设定决策树数量为 200 棵;MLP 采用双隐藏层结构(64+32 个神经元)与 ReLU 激活函数;KNN 通过交叉验证确定最近邻数 $K=15$;SVR 通过随机搜索优化获得最优参数组合($C=100$, gamma = 'auto'). 在测试集上进行预测,模型性能对比结果如表 2 所示. 其中 XGBoost 模型在多项指标上表现突出:其决定系 R^2 达到 0.92,拟合效果显著优于其他模型;均方误差(MSE)为 13.51,均方根误差(RMSE)为 3.68,平均绝对误差(MAE)为 1.88,三项误差指标均处于较低水平,反映出该模型在预测精度上的优势,因此选定其作为解释性分析的模型.

表 2　模型性能对比

Table 2　Comparison of model performance

模型	R^2	MSE	RMSE	MAE
XGBoost	0.92	13.51	3.68	1.88
Random Forest	0.89	17.08	4.13	2.31
SVR	0.89	16.35	4.04	2.05
MLP	0.76	38.72	6.22	3.41
KNN	0.72	38.63	6.22	3.20

2.2　基于 SHAP 的模型解释性分析

2.2.1　全局解释分析

在最优模型基础上,利用 SHAP 方法对模型进行全局解释分析. 通过计算模型中每个特征变量在所有样本中的 SHAP Value 绝对值的平均值,量化其对科技人才流动预测的贡献度如图 1 所示,横轴为对应特征的 SHAP 值,其绝对值反映影响强度;纵轴为按贡献度降序排列的特征变量;点的颜色映射特征值高低,红色表示特征值较高,蓝色表示特征值较低.

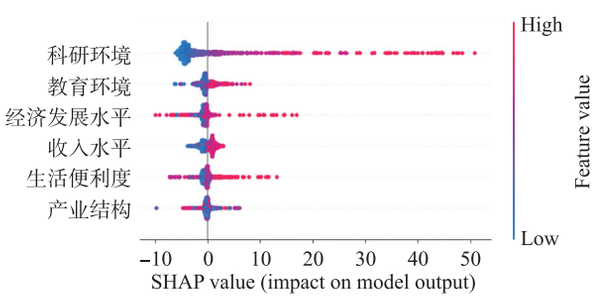


图 1　全局 SHAP 特征概要图

Fig. 1　Global SHAP feature summary plot

从特征重要性来看,科研环境的贡献度显著高于其他特征,表明其对科技人才流动具有主导性影响. 优质科研环境通过提供先进实验设施、充足经费支持、开放学术平台及创新生态等核心要素,直接构成科技人才职业发展的基础支撑,从而成为跨区域流动的关键驱动力. 其余特征按贡献度降序排列依次为教育环境、经济发展水平、收入水平、生活便利度和产业结构,均表现出实质性影响. 相比之下,产业结构的贡献度最低,这与产业转型周期长等特性相符.

从特征影响方向来看,科研环境的 SHAP 值分布以正值为主,且以高特征值为主,表明科研环境越优越,越能吸引科技人才流入. 教育环境的高特征值虽然多对应正 SHAP 值,但整体分布中正负值共存,说明教育环境对科技人才流动的作用可能受其他因素调节. 经济发展水平的 SHAP 值有正有负,且分布相对分散,高、低特征值均对模型输出产生影响,表现出异质性影响,欠发达地区受益于经济增量吸引,而发达地区可能受成本挤压产生人才挤出. 收入水平的影响方向相对明确,低特征值多对应负 SHAP 值,表明收入水平低会抑制科技人才流入,而高收入虽不一定是决定性因素,但却是保障科技人才基本需求与提升流动意愿的重要基础. 生活便利度的高特征值多分布在正 SHAP 值区域,但点群较为分散,说明其影响需结合其他条件共同作用,在生活便利度高的区域,若同时具备优质的科研或经济环境,其对科技人才的吸引力会显著增强;反之,若缺乏核心发展要素,单纯的生活便利可能不足以驱动科技人才流动. 产业结构的 SHAP 值正负值交织且特征值高低与 SHAP 值方向无明显关联,反映了其影响的复杂性,不同产业类型对人才的需求存在结构性差异,且产业转型阶段可能会改变科技人才流动的驱动因素.

2.2.2　特征依赖分析

为进一步探究单一特征变量取值对科技人才流动的边际效应,开展 SHAP 特征依赖分析. 该方法将 SHAP 值视为特征作用力的方向性指标,直观反映了各宏观因素对科技人才流动的推动或抑制作用,其绝对值表征作用强度. 如图 2(a)所示,随着科研环境取值的增加,其 SHAP 值呈现明显的上升趋势,表明科研环境越优越,对科技人才流动的促进作用越强. 这一结果印证了科研环境作为科技人才职业发展基础支撑的关键

地位,优质的科研平台不仅能提供先进的实验条件和研究资源,还能通过学术交流、项目合作等渠道拓展科技人才的发展空间,因此成为吸引科技人才流入的关键驱动力.

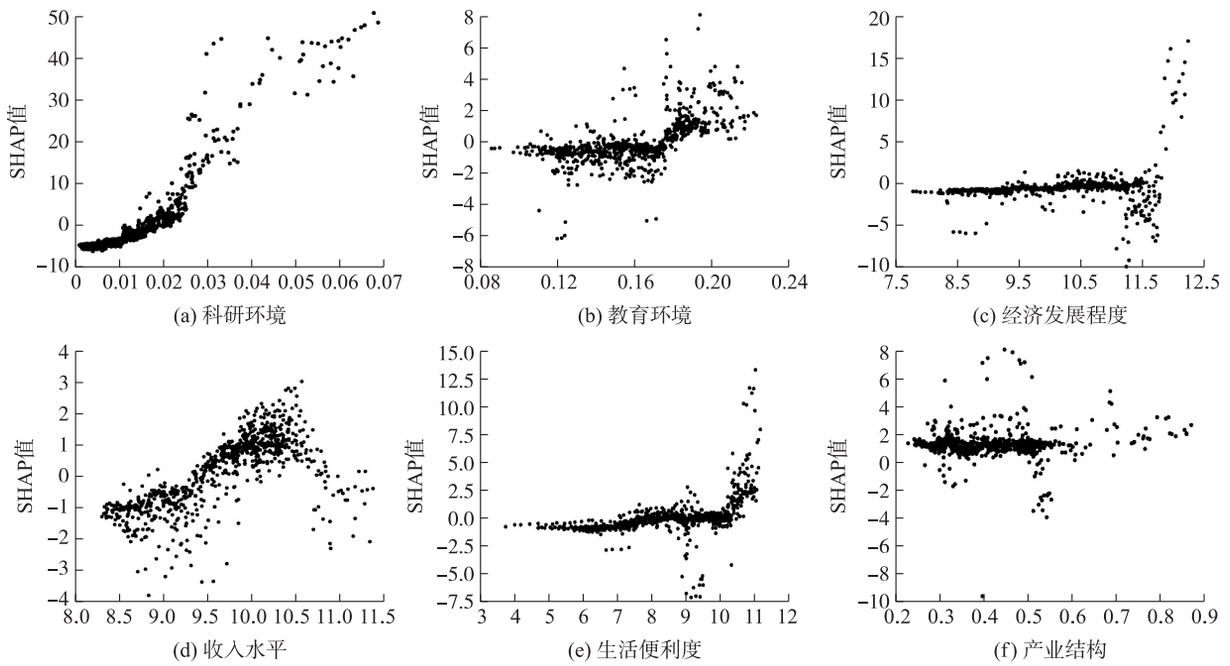


图 2　SHAP 特征依赖图

Fig. 2　SHAP feature dependency plot

与科研环境的明确正向关联不同,教育环境的影响呈现出更复杂的非线性特征. 从图 2(b)可以发现,其 SHAP 值在部分区间存在波动,整体分布不集中,说明教育环境的作用可能存在门槛效应:当教育环境达到一定质量水平后,才能通过人才培养、知识溢出等渠道显著吸引科技人才流入;而在教育环境较为薄弱的区域,单纯的教育投入可能不足以驱动科技人才流动. 这种非线性关系提示,教育环境的影响需结合区域教育资源的实际配置和利用效率来综合判断.

经济发展水平对科技人才流动的影响则呈现阶段性差异. 在图 2(c)中,其 SHAP 值前期较为平稳,并无太大波动起伏,此时经济增长可能主要通过扩大就业机会对科技人才流动产生基础性作用. 但其 SHAP 值在后期产生了较大的波动,说明当经济发展达到一定水平,一方面成熟的经济环境可能通过更完善的产业生态和更多元的发展机会吸引科技人才;另一方面,过度的资源竞争或产业饱和也可能增加科技人才流动的不确定性.

收入水平与科技人才流动的关系呈现出倒 V 型曲线特征. 根据图 2(d),在低收入区间,其 SHAP 值多为负值,表明此时收入优势难以弥补科研配套薄弱、产业支撑不足等短板,对科技人才流动的推动作用有限;在中等收入区间,其 SHAP 值转为正值且集中,说明收入提升成为吸引科技人才流入的关键因素,物质回报的稳定性是此阶段人才决策的核心考量;而在高收入区间,SHAP 值出现回落,这可能与高收入伴随的高生活成本、职场竞争压力有关,收入的边际效益递减效应开始显现,正向影响被部分抵消. 这一结果提示,单纯提高收入并非吸引科技人才的长期策略,需结合其他发展要素协同作用.

生活便利度对科技人才流动的影响存在阈值效应. 从图 2(e)可以看出,在特征值前段,SHAP 值变化较小,表明此时生活配套的不足尚未成为人才流动的主要障碍;但当生活便利度达到一定水平后,SHAP 值明显上升,说明优质的生活环境开始显著增强区域的吸引力,完善的生活配套不仅能提升科技人才的居住体验,还能通过减少通勤时间、提升医疗便利性等间接提高科技人才的职业满意度,从而促进长期定居意愿.

最后,产业结构对科技人才流动的影响较为复杂. 如图 2(f)所示,其 SHAP 值分布较为分散且无明显趋势,表明产业结构的作用可能受区域产业基础与发展阶段的调节. 例如,传统产业主导区域因技术生态落后与职业发展受限,往往抑制科技人才流入;而高新技术产业集聚或处于转型期的区域则凭借创新要素集聚与优质岗位供给形成科技人才引力. 此外,不同产业演进阶段对科技人才技能需求的差异性,进一步导致其与

流动关联方向的分化,因而需结合具体产业特征与发展进程进行动态分析.

2.2.3　特征交互分析

为了进一步捕捉特征变量间的协同或拮抗效应,开展特征交互分析. 具体而言,首先基于 XGBoost 基础模型,利用 SHAP 的 TreeExplainer 计算特征交互值,即特征对之间相互作用的强度;其次,设定交互值阈值为 0.5,筛选出平均交互值绝对值较高的特征对,最终选取“经济发展水平、科研环境”和“科研环境、生活便利度”两组高交互特征(见图 3);最后,通过特征乘积方法生成新交互特征,即“经济发展水平×科研环境”、“科研环境×生活便利度”,并将其加入原特征空间. 重新训练模型的决定系数 R2 提升至 94.5%,相比之前增加了近 3%,均方误差(MSE)为 8.9933,相比之前降低了 4.5,表明特征交互有效优化了模型拟合能力和预测准确性.

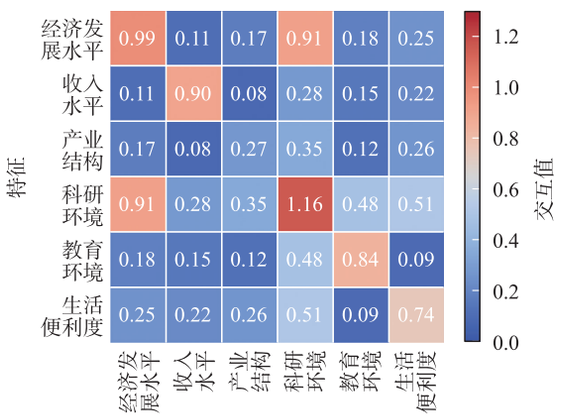


图 3　特征交互热力图

Fig. 3　Feature interaction heatmap

图 4(a)展示了科研环境与生活便利度之间的交互关系. 横轴为科研环境取值,纵轴为 SHAP 交互值,点颜色表示生活便利度水平. 可以发现,当科研环境处于较低水平时,无论生活便利度高低,交互值均接近零,说明生活便利条件难以有效弥补科研环境短板,无法产生显著协同影响. 随着科研环境提升至中等水平,交互值明显转为正向,反映二者开始发挥协同促进作用,共同吸引科技人才流入. 然而,当科研环境进一步达到较高水平时,交互值再次回落,表明在科研环境本身已很强的地区,生活便利度的边际效用趋于饱和,二者的交互效应相应减弱. 图 4(b)呈现了经济发展水平与科研环境之间的交互作用. 横轴为经济发展水平,纵轴为 SHAP 交互值,点颜色代表科研环境水平. 在经济发展水平较低阶段,无论科研环境优劣,交互值均趋于零附近,二者难以形成有效互动. 随着经济发展水平提升,交互效应出现显著分化:在科研环境优良的区域,交互值大幅上升,凸显高水平经济与优质科研环境相结合对吸引科技人才的关键作用;而在科研环境薄弱的地区,交互值则向负区间偏移,表明二者未能形成合力,甚至产生抑制科技人才流动的负面效应. 这一结果反映出,只有当经济发展与科研环境均达到一定阈值并实现良好匹配时,才能充分发挥协同引才效应.

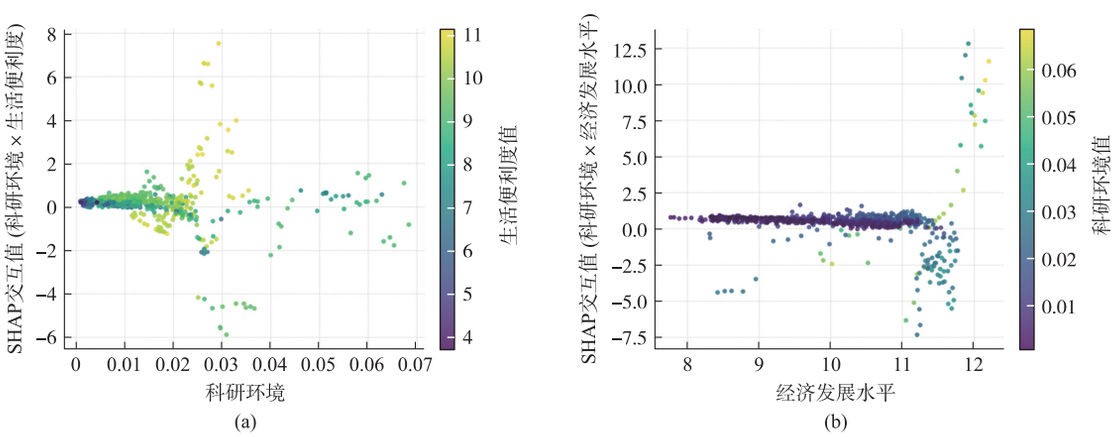


图 4　SHAP 特征交互图

Fig. 4　SHAP feature interaction plot

2.3　分区域解释分析

2.3.1　东部地区

如图 5(a)所示,东部地区的科技人才流动受科研环境与经济发展水平的双重主导. 其中,科研环境 SHAP 值高值区集中于正向区间且分布范围较广,表明东部地区凭借优质的科研平台、丰富的创新资源及活跃的产学研合作生态,持续发挥着创新策源地对科技人才的集聚效应. 经济发展水平的 SHAP 值虽以正向贡献为主,但高特征值区域存在局部负向扰动,这可能与东部核心城市生活成本攀升引发的挤出效应有关,即

部分科技人才为平衡生活压力与职业发展,可能向周边或二线城市迁移,形成核心城市辐射、周边城市承接的人才流动新格局. 此外,产业结构与生活便利度的 SHAP 值分布相对分散且绝对值较低,表明其对科技人才流动的影响力度较弱,这或与东部地区产业结构已进入高技术化、服务化阶段,生活便利度整体较高且区域差异缩小有关.

2.3.2　中部地区

如图 5(b)所示,在中部地区,科研环境、经济发展水平和教育环境对科技人才流动具有关键影响. 其中,科研环境的 SHAP 值高值区虽分布范围较东部明显收窄,但仍呈现显著正向特征,反映中部地区通过持续增加科研投入、深化产学研协同创新,逐步形成了具有区域特色的科研吸引力. 经济发展水平的 SHAP 值分布较为分散,高特征值区域的正向影响表明中部地区经济增速的提升对科技人才具有积极拉动作用,但区域内部经济分化可能导致影响的不确定性. 教育环境的影响方向相对明确,低特征值多对应负向影响,反映教育资源配置不均衡可能抑制人才流入;而高特征值区域集中于正向区间,表明优质教育资源的集聚不仅直接吸引科技人才,还通过培养本地人才为区域发展提供支撑,有效提升了人才自给能力.

2.3.3　西部地区

如图 5(c)所示,西部地区科技人才流动的核心驱动因素为科研环境,同时生活便利度、收入水平和经济发展水平的影响力逐步显现. 其中,科研环境 SHAP 值高值区显著正向,反映西部地区虽整体科研资源相对薄弱,但通过布局重大科技基础设施、建设特色学科平台等,已形成对特定领域科技人才的差异化吸引力. 生活便利度的 SHAP 值分布虽分散,但正向区间的绝对值较高,这可能与西部地区近年来基础设施的持续改善有关,生活便利度的提升虽未成为主导因素,但已成为增强人才留存意愿的重要补充. 收入水平和经济发展水平的 SHAP 值高特征值集中于正向区间,表明西部地区通过提高薪资待遇、扩大经济规模,有效增强了对科技人才的经济引力. 值得注意的是,产业结构的高特征值多对应负向 SHAP 值,可能与西部特色产业的技术密集度较低有关,需通过产业升级进一步提升对科技人才的需求匹配度.

2.3.4　东北地区

如图 5(d)所示,在东北地区科技人才流动中,科研环境具有主导性影响,产业结构和经济发展水平也起到了较为重要的作用. 其中,科研环境 SHAP 值高值区虽存在双向影响,但正向区间的覆盖范围较广,反映东北地区依托哈尔滨工业大学等科研重镇形成的科研实力,仍是吸引科技人才的核心基础,但区域内部科研资源分布不均可能导致影响的波动性. 产业结构的 SHAP 值分布虽分散,但高特征值区域呈现显著正向特征,

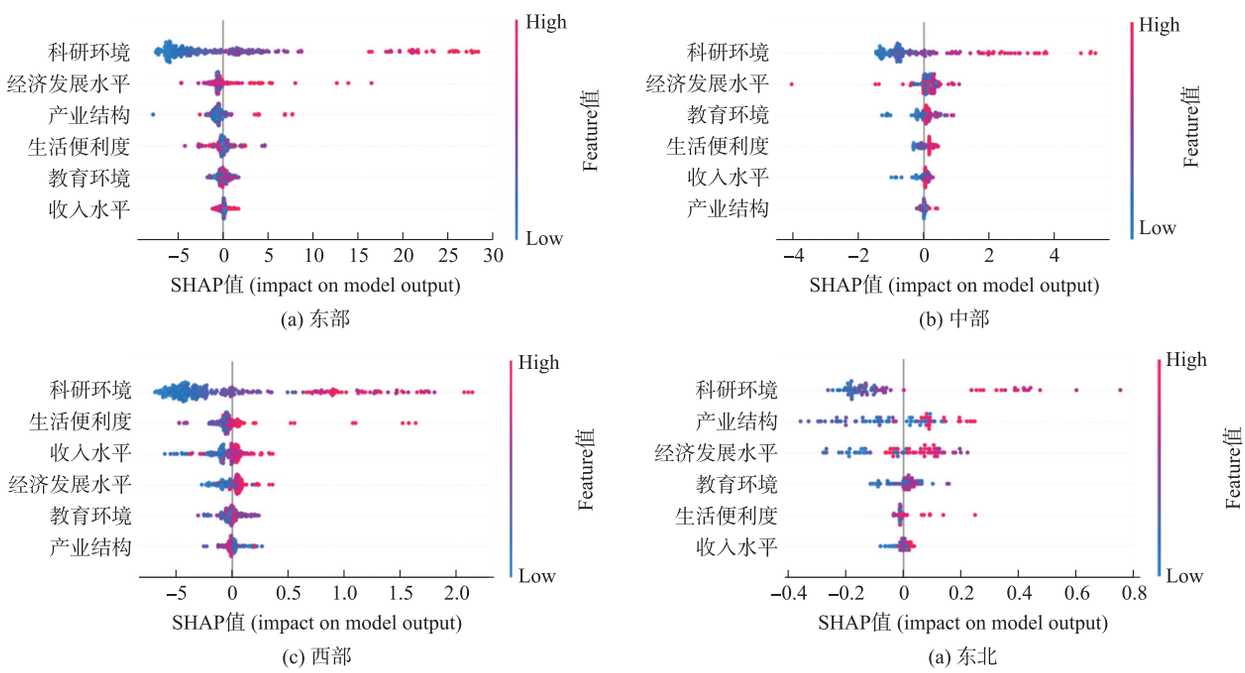


图 5　分区域 SHAP 解释力图

Fig. 5　Regional SHAP force plot

表明东北地区通过装备制造智能化等传统产业升级,以及冰雪经济、生物医药等新兴产业培育,正逐步形成对相关领域科技人才的需求拉动. 经济发展水平的 SHAP 值正负交织,高特征值区域的正向影响提示东北振兴战略推动的经济企稳对科技人才流动有积极作用,但区域内部经济活力不足仍可能导致影响的波动性.

2.4 稳健性检验

本文采用改变数据时序划分的方法来检验结果的稳健性. 首先,将原始随机划分的训练集/测试集(8∶2)调整为严格依时间顺序的"前 N 年训练、后 M 年测试"划分方式(1978—2010 年为训练集,2011—2023 年为测试集). 进一步地,引入滚动窗口验证(窗口长度 10 年,步长 1 年)进行补充检验. 对比不同划分方式下的模型性能与特征重要性排序发现,XGBoost 模型均保持最优预测精度,且科研环境、教育环境等核心因素的 SHAP 全局重要性排序高度一致,表明研究结果具有稳健性.

3 结论

本文基于 1978—2023 年中国 31 个省(直辖市、自治区)的面板数据,综合运用 XGBoost 机器学习模型与 SHAP 可解释性方法,深入剖析了影响科技人才流动的重要因素及其作用方式,得出以下结论:(1)从影响程度来看,各宏观因素对科技人才流动的影响力由大到小依次为科研环境、教育环境、经济发展水平、收入水平、生活便利度和产业结构. (2)从特征变量具体取值来看,科研环境与科技人才流动呈现正向线性关联,而其余因素则表现出复杂非线性关系. (3)从特征交互作用来看,经济发展水平与科研环境、科研环境与生活便利度之间存在显著协同效应. (4)从区域异质性来看,东部、中部、西部及东北地区的科技人才流动的主导因素存在显著区域差异.

基于上述研究结论,本文提出如下建议:

第一,根据宏观影响因素的主次排序,政策制定应优先聚焦核心驱动要素并强化协同. 科研环境作为首要驱动力,应确立为全域性战略支点,通过扩容重大科技基础设施、深化产学研协同机制、构建开放创新生态,强化对科技人才职业发展的支撑效能;同步提升教育环境能级与收入水平的区域竞争力,确保其在吸引和留住人才中发挥关键作用;对于经济发展水平与生活便利度,应侧重其与核心要素的配套协同,将经济增量有效转化为对科研、教育的持续投入,并通过公共服务提质增效提升和基础设施互联互通,增强对科技人才的综合吸引力. 产业结构优化则需立足长远,通过前瞻性产业规划和政策引导,逐步提升区域产业的技术密集度与创新活力,培育更适配科技人才发展的岗位生态.

第二,针对影响因素的复杂非线性特征,政策制定需提升精准性与动态适配能力. 教育环境的影响需通过深化教育评价改革,避免低效教育投入;经济发展水平的影响需区分阶段施策,欠发达地区应释放经济增量对人才的吸附效应,发达地区则需重点应对资源过度集中导致的流动波动;收入水平政策应把握其倒 V 型曲线规律,在中等收入区间强化物质回报的稳定性预期,在高收入区域则需转向科研自主权、成果转化收益等非货币激励,以抵消收入边际效益的递减;生活便利度的提升需分层推进,未达标区域优先补足基础性民生短板,已达标区域应转向智慧城市、生态宜居等品质升级. 产业结构的影响需结合区域基础动态调整,传统产业主导区域可通过智能化、绿色化转型催生新技术人才需求,新兴产业集聚区域则需强化创新要素与岗位体系的匹配度,避免因结构性错配引发的人才无序流动或短缺.

第三,基于特征交互分析揭示的关键协同机制,政策制定应系统构建宏观影响因素间的联动路径,形成多要素高效耦合的引才体系. 重点推动科研环境与经济发展水平的战略协同,在区域发展规划中确保二者的同步优化与动态匹配,尤其在经济发展已达较高水平的地区,需相应加强科研平台建设与研发投入,防范因科研环境滞后导致的区域人才吸引力结构性弱化;同时,在科研基础设施布局中系统考量生活便利性要素,精准配置高品质公共服务、便捷交通与生态设施,充分发挥二者在人才吸引中的协同增益效应. 对于科研环境相对成熟的地区,则应转向构建更多元的非经济激励生态,通过体制机制创新持续释放科研活力. 最终,形成以科研环境为核心、多要素系统联动的区域人才治理新格局.

第四,各区域应立足自身资源禀赋与核心驱动力,制定差异化引才聚才策略. 东部地区需依托科研环境与经济发展的双主导优势,构建核心城市与周边区域的协同机制,同时缓解高生活成本对人才的挤出效应,促进人才在区域内优化配置与高效流动. 中部地区应以建设区域性科研高地为锚点,提升高端人才集聚与自主培养能力;同时积极承接产业转移、培育本土龙头企业以扩大经济规模效应,并大力推动教育资

源优质均衡发展,增强区域人才吸引力与根植性. 西部地区需坚持以科研环境为核心驱动,通过布局特色优势科研平台与重大项目牵引高端人才流入,同步构建高便捷性生活圈与完善服务配套体系以提升人才留存率. 东北地区则需充分激活其科研积累与产业基础潜能,聚焦装备制造智能化、冰雪经济等特色领域定向引才育才,同时依托东北振兴战略促进经济企稳回升,全面提升区域人才竞争力.

[参考文献]

[1] 黄海刚,曲越. 孔雀东南飞:经济转型与精英科学家流动[J]. 华中科技大学学报(社会科学版),2019,33(3):63-72.

[2] 柳瑛,苏丽锋,史薇. 中国科技领军人才流动的影响要素研究[J]. 技术经济,2025,44(4):25-35.

[3] Zha Q. What factors influence the direction of global brain circulation:the case of Chinese Canada Research Chairholders[J]. Compare:A Journal of Comparative and International Education,2016,46(2):214-234.

[4] YUE M,LI R,OU G,et al. An exploration on the mobility of leading research talents in China:from the perspective of distinguished young scholars[J]. Scientometrics,2020,125(2):1559-1574.

[5] 郭俊华,徐倪妮. 高新技术企业科技人才流动影响因素研究:以上海市为例[J]. 数量经济研究,2020,11(4):90-104.

[6] 徐娟,王建平. 中国大学高层次青年人才流动的影响因素:基于5类项目人才履历追踪的实证研究[J]. 现代大学教育,2021,37(3):78-87.

[7] 朱英,郑晓齐,章琰. 中国科技创新人才的流动规律分析:基于国家"万人计划"科技创新领军人才的实证研究[J]. 中国科技论坛,2020(3):166-173.

[8] 刘杨,有书涵,周建中. 我国科技智库人才流动特征及影响因素定量研究[J]. 科学学研究,2025,43(8):1588-1599.

[9] 王腾,关忠诚,郑海军,等. 全球顶尖科技人才流动的影响因素探析:基于扎根理论的分析[J]. 科学学研究,2024,42(7):1376-1386.

[10] 王冲. 基于灰色关联分析的高校科技人才流动影响因素及策略分析:以吉林省高校为例[J]. 情报科学,2019,37(5):47-52.

[11] 柳美君,李伟平,杨斯杰,等. 长三角区域一体化政策对科技人才流动的影响研究[J]. 科学学研究,2024,42(4):733-745+862.

[12] 王焘,马亚雪,吴柯烨. 区域新兴产业创新生态系统联动效应对科技人才流动影响研究[J]. 现代情报,2024,44(11):139-151.

[13] 陈茜,马向平,贾承丰,等. 基于决策树ID3算法的人才留汉吸引政策研究[J]. 武汉理工大学学报(信息与管理工程版),2019,41(2):148-153.

[14] Al-Quhfa H,Mothana A,Aljbri A,et al. Enhancing talent recruitment in business intelligence systems:a comparative analysis of machine learning models[J]. Analytics(basel),2024,3(3),297-317.

[15] Gazi M S,Nasiruddin M,Dutta S,et al. Employee attrition prediction in the USA:a machine learning approach for HR analytics and talent retention strategies[J]. Journal of Business & Management Studies,2024,6(3),47-59.

[16] Chen T,Guestrin C. Xgboost:a scalable tree boosting system[C]//Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. New York,USA,2016:785-794.

[17] Lundberg S M,Lee S I. A unified approach to interpreting model predictions[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York,USA,2017:4768-4777.

[18] 李婧,产海兰. 空间相关视角下R&D人员流动对区域创新绩效的影响[J]. 管理学报,2018,15(3):399-409.

[19] 刘诗昆,姚奕. 城市科技人才流动水平的测算及其对绿色全要素生产率的影响研究:基于空间回归模型[J]. 南京师大学报(自然科学版),2025,48(2):1-9.

[20] 窦旺胜,秦波,郝美竹,等. 中国青年流动人口城市选择的空间特征及影响因素[J]. 地理研究,2023,42(5):1234-1247.

[21] 徐倪妮,郭俊华. 科技人才流动的宏观影响因素研究[J]. 科学学研究,2019,37(3):414-421,461.

[22] 齐宏纲,产健,师俊杰,等. 中国境外高水平科技人才回流的空间格局及影响因素[J]. 热带地理,2025,45(2):223-237.

[责任编辑:陆炳新]