何为企业数字技术产出的创新鸿沟？——基于多种可解释模型的比较分析

刘佳仪[1]，陈丽宇[2]，黄颖[1]，程从华[1*]
1. 肇庆学院 数学与统计学院，广东 肇庆 526000
2. 肇庆学院 经济与管理学院，广东 肇庆 526000
DOI:10.61369/ASDS.2026100009

摘　　要： 中国企业数字技术创新产出在数字经济不断推进的大环境下呈现出很明显的“零膨胀”和“长尾”双重分布状态，也就是很多企业长期没有专利产出，少数企业却占据了大部分创新成果，公司治理结构在这里面也起到一定作用，本文把这种“创新鸿沟”拆成“门槛鸿沟”和“差异鸿沟”两个前后递进的层次，使用了2000到2024年中国A股上市公司45848条年度数据，搭了一个四层递进式模型框架，第一层用可解释模型打底，第二层用黑箱集成模型做主分析，第三层用SHAP做事后解释，第四层用滚动窗口完成动态验证，本文系统考察了线性回归、Lasso、梯度提升、随机森林、XGBoost、LightGBM这六类模型在样本内外的拟合好坏和误差指标，还通过10折交叉验证确认了结果是否稳健，之后本文结合EBM分类模型的形函数和SHAP依赖图，捕捉变量之间的非线性影响和交互效应，最后本文以5年滚动窗口画出2011到2021年变量重要性变化图，结果显示：（1）推动力量已经从“政策输血”全面转向“内部人力造血”，政府补助的重要性排名从第2位下降到第14位，研发人员数量则上升到首位；（2）监管层总人数有明显的门槛作用，超过临界点后，它会和研发投入形成正向的“放大器”协同；（3）机构投资持股比例呈倒U型，和研发人员数量之间有一个“最优配合区间”；（4）第一股东持股比例、高管货币激励、两职合一、IT背景标签这些传统治理变量持续失效。本文的建模框架为处理具有门槛作用和企业之间差异很大的创新数据提供了可以重复使用的方法范例，实证结论也为优化公司治理结构和精准设计科技政策提供了数据驱动的决策参考。

关 键 词： 数字技术创新；创新鸿沟；可解释模型；多模型比较

What Constitutes the Innovation Gap in Digital Technology Output of Enterprises? A Comparative Analysis Based on Multiple Interpretable Models

Liu Jiayi[1], Chen Liyu[2], Huang Ying[1], Cheng Conghua[1*]
1.School of Mathematics and Statistics, Zhaoqing University, Zhaoqing, Guangdong 526000
2.School of Economics and Management, Zhaoqing University, Zhaoqing, Guangdong 526000

Abstract: The innovative output of Chinese enterprises in the digital technology sector exhibits a distinct dual distribution pattern of "zero expansion" and "long-tail" within the broader context of the rapidly advancing digital economy—namely, many enterprises have historically produced no patent outputs, while a small number of enterprises account for the majority of innovative achievements; corporate governance structures also play a role in this phenomenon. This paper decomposes this "innovation gap" into two sequentially progressive layers: the "threshold gap" and the "differential gap." Utilizing 45,848 annual data points from Chinese A-share listed companies spanning the period 2000–2024, the study constructs a four-tiered progressive model framework. The first layer employs an interpretable model as the foundation; the second layer uses a black-box integration model for primary analysis; the third layer applies SHAP for post-hoc interpretation; and the fourth layer employs a rolling window approach for dynamic validation. This paper systematically examines the fit quality and error metrics of six types of models—linear regression, Lasso, gradient boosting, random forest, XGBoost, and LightGBM—both within and outside the sample, and confirms the robustness of the

基金项目：本研究得到广东省自然科学基金项目（编号：2024A1515010983）；广东省研究生教育创新计划项目（编号:2025JGXM_181）；肇庆学院专业学位研究生课程案例库项目（No.YBANLK2025_09）；肇庆学院第四批创新科研团队项目（编号:TD202410）的资助。
作者简介：
刘佳仪，肇庆学院数学与统计学院，硕士研究生在读，研究方向为大数据分析；
陈丽宇，肇庆学院经济与管理学院，硕士研究生在读，研究方向为数字贸易；
黄颖，肇庆学院数学与统计学院，本科在读；
通讯作者：程从华，肇庆学院数学与统计学院，硕士生导师，研究方向为生存分析。

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

findings through 10-fold cross-validation. Subsequently, by leveraging the shape function of the EBM classification model and the SHAP dependency graph, the study captures nonlinear influences and interaction effects among variables. Finally, using a five-year rolling window, the paper plots the evolution of variable importance from 2011 to 2021. The results indicate that: (1) The driving force has fully shifted from "policy-driven external support" to "internal human capital development"; the importance ranking of government subsidies has declined from second place to 14th place, while the number of R&D personnel has risen to the top position; (2) The total number of regulatory personnel exerts a distinct threshold effect; once this threshold is exceeded, it interacts in a positive "amplification" manner with R&D investment; (3) The institutional investment shareholding ratio follows an inverted U-shaped curve, with an "optimal coordination range" existing between this ratio and the number of R&D personnel; (4) Traditional corporate governance variables—such as the shareholding ratio of the largest shareholder, monetary incentives for senior executives, the integration of dual roles, and IT background indicators—continue to prove ineffective. This paper's modeling framework provides a reusable methodological example for analyzing innovation-related data characterized by threshold effects and significant heterogeneity across firms; its empirical findings also offer a data-driven decision-making reference for optimizing corporate governance structures and precisely designing science and technology policies.

Keywords : **digital technology innovation; innovation gap; interpretable models; multi-model comparison**

引言

在数字经济时代下，数字技术创新能力已经成为企业核心竞争力的一个重要方面，国家“十五五”规划纲要明确要加快数字中国建设，推动数字技术和实体经济深度融合，这样一来数字技术产出的重要性就被提升到了国家战略层面，但是在实际经济活动里，不同企业的数字技术产出分化很明显，有的企业依靠持续的研发投入和较好的治理机制实现了技术上的跃升，还有大量企业却一直处在创新产出低水平徘徊甚至完全没有产出的状态，这种企业间的“创新鸿沟”现象，已经成为制约数字经济高质量发展的一个主要障碍，本文将创新鸿沟分成两个前后递进的层次，第一个层次是“门槛鸿沟”，也就是企业能不能实现“从0到1”的数字技术突破，第二个层次是“差异鸿沟”，也就是企业跨过门槛之后，在创新产出的数量和质量上继续出现分化，这两个层次共同构成了企业数字技术产出差异的结构性根源，所以有必要从公司内部治理的角度去深入分析它的形成原因和破解路径。

针对技术创新驱动因素，现有文献从人力资本、公司治理、政策支持等多个角度进行了讨论，大多数研究采用的是传统线性回归模型，对于变量之间复杂的非线性关系缺乏有效描述，股权集中程度、研发人员占比、薪酬差距等因素的非线性作用形式关注不足，同时不同所有制企业创新驱动路径的系统比较和动态演变趋势同样缺少考察；因此本文选取2000—2024年中国A股上市公司作为研究对象，综合使用内部解释模型和外部解释方法SHAP，对公司结构因素驱动发明专利产出的机制以及其中的非线性特征进行系统考察，并揭示企业数字技术产出差异背后的结构性原因，并为优化公司治理结构、完善科技政策设计提供理论依据和实证参考。

已有研究指出，企业开展数字化转型能够明显促进技术创新产出[1]，并且可以通过企业创新渠道推动高质量发展[2]；数字技术应用可以优化资源配置效率、降低信息不对称程度，还能增强技术吸收和知识整合能力，从而提升创新的数量和质量[3]；进一步来看，数字技术应用在技术溢出影响技术创新产出的过程中发挥调节作用[4]。此外，数字化转型可能通过多个中介机制影响企业财务绩效，这种影响逻辑和创新产出效应需要进行区分[5]，政府补助在这一过程中也发挥重要的调节作用[6]。但是现有研究大多关注数字技术的平均效应，对公司治理机制如何影响数字技术创新的效率和方向缺少深入讨论。

“创新鸿沟”这一概念是从数字鸿沟理论延伸出来的，是指企业在创新产出和能力方面出现的结构性分化[7]；按照“接入—应用—转化”的分析框架，企业在数字技术获取、使用以及价值转化这些环节上存在鸿沟，这些鸿沟会造成创新能力的非连续性差距[8]，而组织惯性、路径依赖以及动态能力方面的不同则是导致这一鸿沟的重要原因[9]；现有研究大多集中在数字技术接入层面的不平等问题上，对于企业内部结构性因素如何塑造创新鸿沟缺少系统检验。

数字技术产出和创新鸿沟之间存在的双向关联，数字金融发展能够通过缓解融资约束来缩小企业间的创新差距，但是这个过程会受

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

到数字鸿沟的削弱效应影响[10]；科技金融政策和数字并购则可以帮助企业跨越创新能力门槛，进而实现技术赶超[11,12]。

已有文献从数字经济政策、公共数据开放以及数字技术应用等视角进行了研究，结果发现数字经济政策可以促进传统企业进行跨界数字创新[13]，公共数据开放能够带来技术创新效应[14]，数字技术应用则有助于企业开展合作创新[15]；总体来看，现有研究仍然大多关注平均水平效应，对于非线性作用、门槛效应以及不同所有制情况下的差异化表现还缺少足够的关注。

基于上述文献综述，可以发现现有研究存在以下不足：一是多依赖线性模型，难以刻画变量间的非线性关系；二是对股权集中度、研发人员占比等因素的非线性作用形态关注不足；三是缺乏对不同所有制企业创新驱动路径的系统比较。本文拟采用多模型对比与可解释框架弥补上述研究缺口，并参考吴学福等人对零膨胀计数型数据进行的 Hurdle 模型的分析框架[16]，对当前问题分两阶段考察创新鸿沟：第一阶段聚焦“门槛鸿沟”，采用 EBM 内在可解释模型识别企业跨越数字技术创新门槛的关键结构条件；第二阶段聚焦“差异鸿沟”，综合选取线性模型、集成模型、树模型三类模型，各取两个代表模型进行评估指标的比较，选取最佳模型，结合 SHAP 事后解释方法比较各方法来揭示跨越门槛后企业间创新产出差异的驱动机制。

基于上述研究设计与实证分析，本文的创新点主要体现在以下三个方面；首先，在研究视角上，本文突破了传统线性回归模型只关注平均效应的局限，针对企业创新数据普遍存在的“零膨胀”与“长尾分布”特征，构建了分阶段建模框架；围绕创新鸿沟的两个核心问题，一个是企业如何跨越数字技术创新门槛，另一个是企业间创新产出差异的驱动因素，本文分别考察了从“无专利”到“有专利”和从“低产出”到“高产出”这两个阶段的驱动逻辑，揭示了其中明显的异质性。

其次，在模型构建上，本文构建了“内在可解释模型基准参照—黑箱集成模型主分析—SHAP 事后解释机制挖掘”的递进式建模管线；该管线先利用 EBM 二分类功能来识别跨越创新门槛的关键条件，再通过三类六种模型的纵向比较确定最优预测模型，接着借助 SHAP 依赖图捕捉变量之间的非线性形态和二阶交互效应，最后采用时间滚动窗口来检验模型结论的动态稳健性；这一设计从数据结构特征出发逆向推导模型需求，为处理具有门槛效应和强异质性的企业创新数据提供了可以重复使用的分析范式。

再次，在动态分析上，本文首次把时间滚动窗口和 SHAP 重要性排序结合起来，绘制了2011—2021年变量重要性动态演变图谱，从时间顺序角度考察了各个解释变量解释力的变化；这种方法不仅丰富了创新驱动因素的动态证据，也为统计建模中模型结论的时间稳定性问题提供了一个直接的实证检验方案。

一、变量选取与建模数据基础

（一）数据来源

本文数据来源于国泰安（CSMAR）数据库，包含2000-2024年中国 A 股上市公司面板数据和高管数据。合并清晰过程中以“股票代码—年度”为匹配对象。样本筛选过程中，剔除了金融类上市公司、ST 及 *ST 类企业以及主要变量存在严重缺失的观测样本，最终用于本文建模分析的数据集包含4887个企业共45848条年度观测值。以上数据皆公开，数据来源真实、可验证，能够有效支撑本文的实证分析。

（二）特征选取

为系统考察公司结构背景对数字技术产出的影响机制，本文分别从以下维度选取变量：企业特征、财务状况、成长能力、创新投入、行业环境、股权结构、公司治理、管理层特征、创新人才、政策扶持及数字化转型。选取的变量分别作为控制变量与解释变量。

1. 被解释变量

被解释变量采用上市公司‘发明专利授权量’，其记录了企业当年发明专利授权量的个数，能够体现企业数字技术产出量，因此使用该指标作为被解释变量用于研究企业之间产生差异的鸿沟何在。‘专利产出’是通过‘发明专利授权量’构造的指标，是一个二分类指标，0表示企业在当年无产出，1则有，这一指标体现了数字技术产出的门槛，即企业如何从无到有。

2. 解释变量

解释变量从以下五个方面选取：第一，股权结构方面选取第一股东持股比衡量股权集中度、股权性质、机构投资持股比，以考察股权结构在公司结构背景系统下的作用；第二，公司治理方面选取董事会规模、独立董事人数、监事会规模、监管层总人数及持股数量、两职合一，以考察治理结构对创新决策的影响；第三，管理层特征方面选取高管薪酬总额与高管前三薪酬总额、高管教育背景与信息技术背景，这个方面用于考察管理团队的知识资本与技术经验；第四，创新人才方面选取研发人员数量衡量绝对投入规模、研发人员占比考察团队结构配置的边际影响；第五，政策扶持方面选取政府补助金额衡量政策支持力度。

3. 控制变量

为缓解遗漏变量偏误，选取企业年龄与企业规模控制企业特征，选取总资产净利润率、资产负债率、固定资产占比控制财务状况，选取托宾 Q 值与营业收入增长率控制成长能力，选取研发投入总额与强度控制创新投入水平，引入行业集中度控制行业竞争格局。

二、数据清洗

为保障后续实证回归结果的稳健性与结论解释力，本文基于原始企业面板数据，遵循”变量统一—缺失修复—异常管控—尺度变换”的逻辑框架，开展多源数据系统性清洗与标准化预处

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

理。本文原始数据取自企业财务报表、公司治理数据库及创新投入专题数据库，以“股票代码—年度”为匹配标识完成多表对接与样本整合。

（一）数据清洗

1. 多源数据整合与主表构建

本文研究涉及财务、公司治理、创新特征多维度数据表，为统一样本统计口径，以企业年度财务数据为主表，选取“股票代码、年份”作为联合主键，对公司治理表、创新特征表依次采用“左连接（Left Join）”方式合并。该方式可最大限度保留主表企业年度观测样本，同步补充治理、创新维度变量信息，规避内连接、交叉连接造成的样本大量损耗问题。

2. 主营业务文本分类编码

原始主营业务为非结构化文本数据，存在行业表述繁杂、口径不统一、描述冗余等问题。本文采用“关键词匹配 + 文本筛选”方法对主营业务进行标准化分类编码，聚焦软件开发、云计算、大数据、人工智能、信息服务、电子商务、集成电路等数字技术业务关键词，构建企业数字业务虚拟变量：企业主营业务涉足数字技术领域赋值为1，未涉及则赋值为0，实现文本信息可量化建模。

3. 股权性质变量统一赋值

赋值编码规则如下：国有企业、央企、地方国资控股等公有制企业赋值为1；民营企业、外资企业及混合所有制非国资控股企业等非公有制企业赋值为0，统一变量内涵与实证解释口径。

4. 缺失值分层初步处理

通过数据信息核验与变量缺失率统计，分层开展缺失值处理：专利数量、研发投入、政府补贴、技术投入等投入计数类变量，遵循“未披露即未发生”的经济逻辑，缺失值统一赋值为0；股权持股比例、持股金额等核心治理变量，缺失将严重影响实证有效性，直接剔除对应年度观测样本，剔除后样本量仍满足计量建模基本要求。

5. 重复值剔除与格式标准化

统一股票代码、年份字段数据格式与字段类型；以“股票代码 + 年份”为唯一识别维度开展重复值去重，保留字段信息最完整的观测记录，杜绝重复样本干扰回归系数估计结果。

（二）数据预处理

建模前先完成变量定义与描述性统计，刻画变量分布形态、离散程度及潜在异常特征，在此基础上开展精细化预处理。

1. 缺失值精细插补修复

本文采用“逻辑回归插补法”进行缺失值拟合填补，这能有效捕捉到变量间统计关联关系，从而更精准地估算缺失数值，相比传统均值填补更能保留原始数据分布结构。

2. 异常值识别与缩尾处理

本文利用箱线图同时结合“四分位距”法则，识别与判断异常值，并发现多数连续型变量存在极端的异常值。为避免异常值对参数估计的影响，本文对所有连续变量进行1%–99% 缩尾处理。

3. 连续变量 Z–score 标准化

资产规模、盈利能力、研发强度等连续变量量纲差异显著，为消除量纲干扰、提升模型收敛速度与系数可比性，对全部连续变量进行 Z–score 标准化：

$$Z = \frac{X - \mu}{\sigma}$$

式中：μ 为变量均值，σ 为变量标准差。标准化后变量满足均值为0、标准差为1。

4. 计数型变量对数变换

专利申请量、政府补贴次数、投资事件数等计数变量呈明显右偏长尾分布，本文对非零观测值进行自然对数变换：通过对数变换压缩长尾、弱化极端大值干扰，改善变量正态性；取值为0的样本保留原值，坚守变量经济含义不变。

经过上述数据预处理操作后，最终得到口径统一、结构完整、统计特征稳定的企业面板数据集。预处理显著优化了数据的完整性、一致性与分布合理性，为后续基准回归、模型比较、稳健性检验及拓展分析提供坚实的数据支撑。

（三）变量说明及描述性统计

表1与表2分别报告了控制变量、被解释变量与解释变量的描述性统计结果。被解释变量‘发明专利授权量’的均值为13.90，但中位数为0，且标准差高达153.05，表明大量企业都没有发明专利产出，且有产出的企业之间差距较大，呈现严重的“零膨胀”与“长尾”现象。若此时采用传统回归方法可能出现严重的过拟合，因此进行两阶段建模与多模型比较是有必要的。另外发明专利申请量的均值为89.12、标准差为786.22，进一步验证了创新产出的极端不均衡性。

在解释变量方面，研发人员数量的均值为310.65，但中位数为0，说明大多企业尚未配备专职研发人员，与发明专利授权量呈现相似的现象。这一现象预示着企业在结构治理上跨越创新门槛很可能收到研发人员规模的影响。监管层总人数的均值为16.49、中位数为16、标准差为3.96，分布相对集中，但其对创新的影响是否呈现非线性形态有待模型验证。机构投资持股比的均值为45.56%、标准差为24.52%，覆盖从极低到极高持股的完整区间。政府补助的均值与中位数差异悬殊，表明补助资金高度集中于少数企业。

表1：解释变量与被解释变量的定义和描述性统计

变量名	定义	样本量	均值	中位数	标准差
解释变量					
第一股东持股比	第一大股东持股数量 / 总股数，衡量股权集中度与大股东控制程度	43157	34.9487	32.55	15.3895
股权性质	虚拟变量，0表示非公有制，1表示公有制，控制产权性质差异股权性质。	45848	0.3785	0	0.485
董事会规模	董事会成员总人数，衡量董事会规模与治理结构	45848	8.7308	9	1.9057

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

独立董事人数	独立董事人数，衡量董事会独立性与监督效率	45848	3.1667	3	0.7284
研发人员数量	研发人员总人数，衡量创新人力资源投入	45848	310.6479	0	1753.455
研发人员占比	研发人员 / 员工总数，衡量创新人才结构	45848	7.9729	0	12.7944
监事会规模	监事会成员人数，衡量内部监督机制强度	45848	3.6574	3	1.2427
高管薪酬总额	高管年度薪酬总额，衡量管理层激励水平	45848	3936863.21	2520750	5626588.61
政府补助	当年获得的政府补助金额，衡量政策支持力度	45848	23408090.33	1091207.56	291331341
机构投资持股比	机构投资者持股比例，反映机构投资者对公司的关注和认可程度	45384	45.5592	47.7233	24.5245
监管层总人数	公司监管层的总人数，反映公司治理结构的规模	45848	16.485	16	3.9593
监管层持股数量	公司监管人持有的公司股份总数	44199	49836993.82	689931	127085381.6
高管前三薪酬总额	公司年度内薪酬最高的三位高级管理人员的薪酬总和（包括工资、奖金、福利等）	45848	2221881.13	1541000	2858113.65
两职合一	董事长与总经理是否为同一人；1= 是，0= 否，衡量治理结构	45848	0.0822	0	0.2747
高管教育背景	董监高是否具有研究生教育背景；1= 是，0= 否	45848	0.2914	0	0.4544
主营业务	主营业务数字化转型程度，是否有数字技术类业务，0表示不是，1表示是	45848	0.1919	0	0.3938
高管信息技术背景	虚拟变量，CEO 具有技术从业背景时为1，否则为0	45848	0.0384	0	0.19
被解释变量					
发明专利授权量	发明专利授权数量，衡量高质量创新产出	45848	13.9036	0	153.0518
专利产出	企业是否有专利产出，0表示不是，1表示是	45848	0.3364	0	0.4725

控制变量大多呈现非正态分布特征，总资产净利润率的均值为0.56%，标准差109.86%、托宾 Q 值 D 与营业收入增长率的均值与中位数之间同样存在较大偏离，同样暗示我们样本内存在极端盈利差异。综合数据特征，本研究在模型选取上需要能够处理“有 / 无”二分类门槛问题的分类模型、能够捕捉非线性关系与高阶交互作用的回归模型、具备可解释性以揭示变量作用方向的参考模型。

表2：控制变量的定义和描述性统计

变量名	定义	样本量	均值	中位数	标准差
企业年龄	企业成立至观测年度的年限，反映企业生命周期与经营成熟度	45843	16.51283	16	6.279152499
总资产	期末总资产账面价值（元），衡量企业整体资产规模	45848	54401166807	2931758243	746476150138
企业规模	总资产的自然对数，消除规模异质性，标准控制变量	45848	22.0193	21.7989	1.5191
总资产净利润率	净利润 / 平均总资产，衡量资产利用效率与整体盈利能力	45848	0.5584	0.0364	109.8558
托宾 Q 值 D	资产重置成本法托宾 Q，衡量企业成长性与投资机会	44133	4.2629	2.1289	243.5202
营业收入增长率	营业收入同比增长率，衡量企业成长能力	44916	-3.7961	0.1011	800.4029
资产负债率	负债总额 / 资产总额，衡量企业长期偿债能力与财务杠杆水平	45848	0.155	0	1.1891
固定资产占比	固定资产净额 / 资产总额，衡量企业资产结构与重资产程度	45848	0.0686	0	0.1388
行业集中度 D’	赫芬达尔指数 D，基于营业收入衡量行业竞争程度	45848	0.0334	0	0.0878
研发投入总额	年度研发投入资金总额，衡量创新投入规模	45848	135552502.5	20741293.02	748212118

三、多模型构建与性能比较

（一）模型说明

1. 模型评估指标与选取

为全面评价各模型的预测性能，本案例采用样本内拟合优度、样本外拟合优度、均方误差、平均绝对误差四个评估指标。样本内拟合优度 R_{in}^2 和样本外拟合优度 R_{out}^2 用于衡量模型对训练数据和测试数据的泛化能力，指标计算方法详见公式（2）和（3）。

$$R_{in}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

$$R_{out}^2 = 1 - \frac{\sum_{i \in Test}(y_i - \hat{y}_i)^2}{\sum_{i \in Test}(y_i - \bar{y}_{Train})^2} \tag{3}$$

均方误差 MSE 可以反映预测值与真实值之间偏差的平方的期望，然而均方误差对异常值较为敏感，为模型预测提供更多稳健性指标参考，本研究还使用了可以反映预测值与真实值之间绝对偏差平均水平的平均绝对误差 MAE。指标计算方法详见公式（4）和（5）。

$$MSE = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{5}$$

另外，本文研究聚焦于公司结构背景数据对数字技术创新产出的影响，需要在保证模型预测效果的同时，清晰识别关键影响变量，因此采用可解释机器学习思路进行模型选取。可解释机器学习（Explainable Machine Learning, XML）的目标是在保证模型预测性能的同时，提高模型决策过程的透明度，根据解释方式不同可分为内在解释和事后解释 [16]。

内在可解释（Intreinsic Interpretability）指模型自身结构简单，使用者能够直观看到模型内部结构。常见内在可解释模型

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

有：线性回归、逻辑回归、决策树等。

事后可解释（Post-hoc Interpretability）指模型训练之后，使用一定方法增强模型的可解释性，挖掘模型学习到的信息。常用事后可解释的方法有：LIME 事后可解释方法、SHAP 事后可解释方法[17]。

图1所示的是可解释机器学习中模型精度和模型可解释性的关系，由香港大学张爱军教授提出。横轴代表模型的可解释性，越往正方向模型可解释性越高；纵轴代表模型的精度，越往正方向模型精度越高。可以看到内在可解释的一些模型精度相对低，但是模型可解释性高，需要使用到事后可解释的模型精度高，可解释性偏弱。

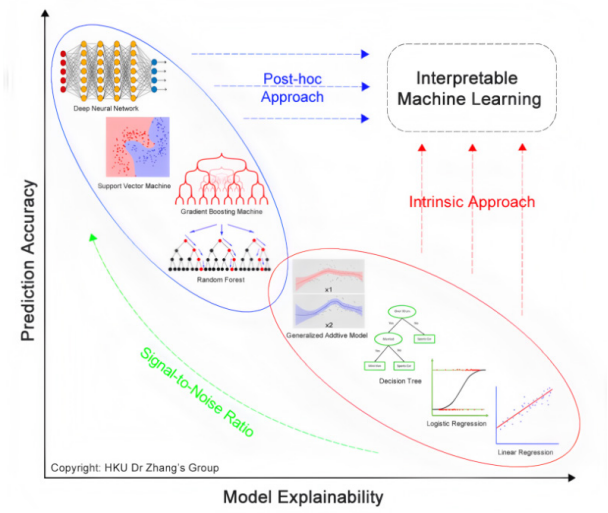


图1：可解释机器学习：模型精度和模型可解释性的关系

鉴于此，本案例首先选取适用于二分类变量预测的内解释模型 EBM 模型，来探究企业跨越数字技术创新门槛的因素。再通过多模型比较的方式确定进行回归预测与事后可解释的模型选取，并以此来探究企业数字技术创新的差异何在。

2.EBM 模型

EBM（可解释提升机，Explainable Boosting Machine），是一种内在可解释机器学习模型，其初衷是在保证预测性能接近复杂黑箱模型的同时，提供全局与局部的透明解释[18]。EBM 结合了广义加性模型和梯度提升树，能够提供特征的单一变量与二阶交互作用的变量重要性排序，适用于二分类目标的预测。

其原理是在给定一个分类任务数据，训练样本为 $(D_i=(x_i,y_i), i=1,\cdots,l)$，其中 (x_i) 为输入样本，(y_i) 为样本的类别值，基础形式为：

$$g\left(E\left(Y|X\right)\right)=\sum_{i=1}^{N}f_i\left(X_i\right) \quad (6)$$

为进一步捕捉特征间的非线性交互效应，EBM 引入二阶交互项，将模型扩展为：

$$g\left(E\left(Y|X\right)\right)=\sum_{i=1}^{N}f_i\left(X_i\right)+\sum_{i,j}f_{ij}\left(X_i,X_j\right) \quad (7)$$

其中 $g(\cdot)$ 为链接函数（回归任务为恒等函数，分类任务为 logit 函数），$f_i(X_i)$ 为特征 X_i 的主效应形函数，由梯度提升树拟合得到；$f_{ij}(X_i,X_j)$ 为特征 X_i 与 X_j 的二阶交互效应形函数。

EBM 的训练先通过梯度提升后向拟合训练所有特征的主效应形函数，再基于模型残差，利用 FAST 算法快速筛选并加入 Top-K 二阶交互项。特征重要性以形函数的标准差衡量：$\mathrm{std}(f_u)=\sqrt{E(f_u^2)}$。

EBM 主要的优缺点见下表。

表3：可解释提升机（EMB）主要优缺点

优点	缺点
1. 精度接近复杂黑箱模型，且具备天然可解释性	1. 仅支持二阶交互，难捕高阶特征
2. 支持全局与局部双重解释	2. 高维特征下计算复杂度上升
3. 自动识别筛选二阶特征交互项，降低计算成本	3. 形函数受限，非线性拟合弱于深度模型
4. 数据预处理成本低，无需复杂归一化分箱操作	4. 相比单一决策树，训练耗时略高

3.Linear 与 lasoo

线性回归（Linear Regression）是最基础的参数化回归模型，其核心原理是假设因变量与自变量之间服从线性关系，通过最小化残差平方和来估计模型参数。其数学表达式为：

$$y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta pxp+\varepsilon$$

其目标函数为：

$$\min_{\beta}\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

该模型的优点在于结构简单、可解释性极强、计算效率高，能够直接量化各变量对因变量的边际影响；但缺点同样明显——对多重共线性高度敏感，当特征数量超过样本量时无法求解，且极易过拟合，拟合能力局限于线性关系，无法捕捉变量间的非线性影响过程。

套索回归（Lasso Regression）则加入了 L1 正则化项，通过压缩部分回归系数至零来实现特征选择与过拟合抑制[19]。其目标函数为：

$$\min_{\beta}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

其中，λ 为正则化强度参数，$\sum_{j=1}^{p}|\hat{a}_i|$ 为 L1 惩罚项。Lasso 能够自动筛选不相关变量，提取关键的变量，极大提升了模型的效率；但同样存在弊端，即当预测变量高度相关时，Lasso 倾向于仅选择其中一个而忽略其它变量，导致系数估计存在一定偏误，因此需要谨慎调参。相较于线性回归，Lasso 能够显著提高模型的稳健性与特征筛选能力，但两者均属于线性框架，无法从根本上解决非线性关系的建模问题。

4. 随机森林与 LightGBM

随机森林（Random Forest）是一种基于装袋法的集成学习模型，其通过自助采样法生成多个训练子集，并随机选取特征子集来构建大量相互独立的决策树，最终通过平均或多数投票的方式输出预测结果。其回归预测表达式为：

$$\hat{y}=\frac{1}{B}\sum_{b=1}^{B}f_b(x)$$

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

其中，B 为决策树的总棵数，$f_b(x)$ 为第 b 棵树的预测值。该模型能够自动处理缺失值与异常值，并且敏锐捕捉到非线性关系和特征交互；再通过随机特征选择有效降低了单棵树的方差，使得模型泛化性能优且不易过拟合。但由于构建多个深度较大的树，模型的内存占用大、预测耗时高、可解释性较差，且在噪声较大的数据上容易产生过拟合风险。

轻量级梯度提升机（LightGBM）则是一种基于梯度提升决策树框架的高效集成模型，采用叶节点按叶子生长策略，结合单边梯度采样和互斥特征捆绑技术，在保持高精度的同时大幅提升训练速度。其通用梯度提升目标函数为：

$$L^{(t)}=\sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)}+f_t(x_i)\right)+\Omega(f_t)$$

其中，$l(\cdot)$ 为损失函数，$\hat{y}_i^{(t-1)}$ 为前一轮的预测值，$f_t(x)$ 为当前轮次拟合的决策树，$\Omega(f_t)$ 为正则化项。相比随机森林，LightGBM 的训练速度快、内存消耗低，能够高效处理大规模高维数据、更快地收敛至更高精度。但对小样本数据比较敏感，噪声数据拟合效果差，需要精细调整学习率与树深度等超参数。

5. 梯度提升 Boosting 与 XGBoost

极端梯度提升（XGBoost）是梯度提升框架下的集大成者，其在传统提升算法的基础上引入了二阶泰勒展开、L1 与 L2 混合正则化、列采样以及并行化节点分裂等创新技术 [20]。其目标函数在每轮迭代中的通用形式为：

$$L^{(t)}=\sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)}+f_t(x_i)\right)+\sum_{k=1}^{t}\Omega(f_k)$$

其中正则化项定义为：

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2+\alpha\sum_{j=1}^{T}|w_j|$$

此处，T 为树的叶子节点个数，w_j 为叶子节点的权重，γ、λ 和 α 分别为控制树复杂度、L2 正则化与 L1 正则化的超参数。在每次分裂时，XGBoost 利用损失函数的一阶导数与二阶导数（即海森矩阵）进行节点划分，其增益计算公式为：

$$\text{Gain}=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]-\gamma$$

其中，G 和 H 分别为一阶和二阶导数的累加和。XGBoost 正则化项的引入能够有效抑制过拟合风险，其二阶导数信息能够让树的生长方向更精准，训练效率高的同时能够自动处理缺失值。但 XGBoost 的缺点在于超参数空间庞大，调参过程较为复杂，对硬件资源的消耗也相对较高。

传统梯度提升机（Boosting）同样采用序列化的前向分步算法，每轮迭代拟合上一轮模型的负梯度方向以降低偏差，其迭代公式为：

$$F_m(x)=F_{m-1}(x)+\nu\cdot\gamma_m\cdot h_m(x)$$

其中，ν 为学习率，γ_m 为当前弱学习器的最优步长，$h_m(x)$ 为拟合残差或负梯度的基学习器。传统 GBM 的优点在于模型灵活、可解释性尚可，且通过逐步提升能够显著提高精度，但其训练过程是串行的形式，而不是并行化处理，这就导致训练时间极长。相较于 XGBoost，传统 GBM 在训练速度、泛化性能和工程实现层面均已全面落后，XGBoost 通过二阶近似与结构化正则化超越了经典 GBM 的性能。

以上六类模型的选取既顺应了数据特征，也结合了复杂度、精度、可解释性的比较标准。第一组作为线性框架的上下界，Linear 和 Lasso 作为线性模型类共同检验线性假设在此数据上是否足够。随机森林和 LightGBM 模型则检验集成策略的差异是否显著影响预测精度。此外 Boosting 和 XGBoost 模型用于检验算法迭代是否带来了可观测的性能跃升。三组对照从线性到树模型到梯度提升递进，保证每一组控制变量一致，通过改变和比较核心算法，从而使得模型比较结论具有清晰的归因路径。

（二）模型训练与评估

1.EBM 模型对‘专利产出’的预测

在前文我们将企业数字技术产出的‘创新鸿沟’划分为‘门槛鸿沟’和‘差异鸿沟’，第一步我们针对‘门槛鸿沟’进行研究，即为什么有些企业始终无法跨越门槛实现数字技术产出？为此，本案例在第一阶段采用 EBM 可解释提升机，针对二分类变量“专利产出”建立预测模型。以“专利产出”为被解释变量，采用与第二阶段完全一致的特征集。模型训练采用 EBM 默认参数，按年份分层抽样划分 90% 训练集与 10% 测试集。

EBM 模型评估结果如表 4 所示。在类别不平衡的情况下，我们更加关注的是分类模型的 PR 曲线下方面积，结果显示 PR_AUC 达到 0.7034，说明在这个情况下模型仍然能够有效预测和区分企业数字技术是否有产出。另外，EBM 模型的准确率为 0.7626，查准率为 0.6541，查全率为 0.5959 均达到标准，验证这个模型在当前数据集下有很好的区分能力。

表 4：EBM 模型评估指标

准确率	查准率	查全率	pr 曲线下面积
0.7626	0.6541	0.5959	0.7034

2. 三类模型对‘发明专利授权量’的回归比较

第二步，我们对‘差异鸿沟’进行研究。由于‘发明专利授权量’是一个计数型变量，对他进行回归预测非常考验模型的性能，因此评价标准也更加苛刻，从而选用更多的模型进行比较和稳健性建议。在此我们采用线性回归、集成模型与树模型对‘发明专利授权量’这一被解释变量进行回归。然后筛选出含有专利产出的企业，与 EBM 模型训练一致，均按照年份划分训练基于测试集，每年抽取 10% 作为测试集，90% 作为训练集进行模型训练，并通过网格搜索的方式进行超参数调优。

（1）解释变量适配度

表 5 给出了三种模型在仅含控制变量的基准模型与加入公司治理解释变量的完整模型下的四个评估指标。通过比较三类基准模型与完整模型，判断加入解释变量在不同模型下之后是否能显著提升模型的拟合优度、降低误差。如果可以则说明我们的解释变量对被解释变量是存在显著影响的。

可以看到相比基准模型，随机森林的样本外 R 方从 0.3766 大

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

幅提升至0.5234，是三个模型中表现最佳的模型。梯度提升的完整模型同样优于基准，但表现不及随机森林模型。线性回归的完整模型与基准模型之间差异极小，表明线性模型几乎无法利用公司治理变量提升预测能力。综合以上结果，随机森林在样本外泛化能力上显著领先，是本研究最合适的预测模型，而线性模型的低 Out_R² 验证了其内在可解释性高但精度低的缺陷。这一结论也说明加入的公司结构解释变量能够提升模型拟合能力，公司结构变量对企业数字技术创新存在显著的影响。

表5：模型评估指标表

Model	In_R²	Out_R²	MSE	MAE
Random Forest—基准	0.9149	0.3766	1.1452	0.7532
Random Forest	0.9327	0.5234	0.8756	0.6641
Boosting—基准	0.3782	0.3336	1.2242	0.7846
Boosting	0.4199	0.3767	1.145	0.7633
Linear Regression—基准	0.2203	0.2136	1.4446	0.907
Linear Regression	0.2625	0.2594	1.3605	0.857

（2）模型比较

由于不同模型适应不同种类的数据，为避免模型差异带来结果的不稳定性，选择三类模型进行横向比较。此外为能够使用上最佳的模型进行预测与可解释，我们选取三类共6个模型：Lasso 对比 Linear、XGBoost 对比 Boosting、LightGBM 对比 Random Forest 进行同类模型之间的纵向比较。通过设置调参空间，选取了六个模型中的最佳模型。表6显示各最佳模型中，XGBoost 相比 Boosting 样本内外拟合优度有显著提升，且 MSE 和 MAE 均有所下降，所以选取 XGBoost 进行后续的可解释；随机森林样本内拟合优度虽然高达0.9327，但样本外拟合优度仅有0.5234，说明随机森林模型在这个数据下很可能存在过拟合现象，反之 LightGBM 虽然拟合优度不如随机森林，但也具有一定的拟合效果，且样本内外拟合优度相差较小，因此综合考虑选择 LightGBM 模型进行后续实证；最后是线性模型，这里采用了 lasso 回归和线性回归进行比较，结果显示，二者的评估指标相差不大，Linear 相比 lasso 具有极小的优势。这是因为线性模型比较难以捕捉各变量之间的交互关系，因此线性模型在高维数据下，相比集成模型和树模型效果大大降级，但线性模型可以呈现一些变量与因变量之间的正负关系，也具有一定的参考价值。

表6：检验模型评估指标

Model	In_R²	Out_R²	MSE	MAE
XGBoost	0.7087	0.4907	0.9355	0.6798
Boosting	0.4199	0.3767	1.145	0.7633
LightGBM	0.5213	0.4288	1.0494	0.7288
Random Forest	0.9327	0.5234	0.8756	0.6641
lasso	0.2621	0.2598	1.3597	0.8569
Linear Regression	0.2625	0.2594	1.3605	0.857

图2横轴表示发明专利授权量的真实值，纵轴表示预测值。以散点图形式展示了六个模型在完整模型设定下对测试集样本的预测值与真实值之间的拟合程度。当数据点分布越靠近对角线，表明拟合程度越好。随机森林与 XGBoost 模型的样本店最为紧密地围绕对角线分布，拟合效果显著优于其余模型；梯度提升与 LightGBM 模型次之，数据点离散程度略高；线性模型的数据点分布的最为分散，且存在大量偏离对角线的样本点。这体现了集成学习方法在处理高维、非线性关系的数据具有明显优势。

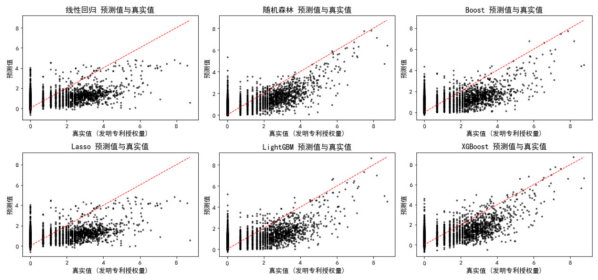


图2：预测值与真实值比较图

（3）10折交叉验证

通过上述模型选取，最终选择模型 Linear、LightGBM、XGBoost 三个不同类别的模型进行后续研究。但可能存在数据集划分不合理导致模型性能不足，为避免这类情况的发生，采用10折交叉验证的方式，进行模型的稳健性检验。

10折交叉验证：将数据随机划分为10组，依次使用每一组作为测试集，其余9组作为训练集，共训练10个模型，得到10个模型的评估指标，并对其取平均值，再与原始划分方法得到的模型评估指标进行比较。

在交叉验证下的结果显示，三个模型的样本内外拟合优度和 MSE、MAE 差距几乎可以忽略，并且文章中使用到的数据划分方式略优于交叉验证时的随机划分。

表7：交叉验证模型评估指标

Model	Sample_In_R2	Sample_Out_R2	MSE	MAE
LightGBM—10fold	0.5207	0.4291	1.0589	0.7351
LightGBM	0.5213	0.4288	1.0494	0.7288
Linear—10fold	0.2624	0.2608	1.3708	0.8650
Linear	0.2625	0.2594	1.3605	0.8570
XGBoost—10fold	0.7062	0.4832	0.9587	0.6874
XGBoost	0.7087	0.4907	0.9355	0.6798

四、构建模型可解释

（一）内在可解释模型

1.EBM 模型探究如何跨越‘门槛’

通过 EBM 模型输出的特征重要性排序显示，影响企业能否跨越数字技术产出门槛的因素众多，本文重点聚焦于四个与公司治理结构密切相关的核心变量：机构投资持股比、监管层总人数、研发人员数量、高管前三薪酬总额。这四个变量在全局重要性排序中名列前茅，说明在控制研发投入等基础因素后，治理结构与人才激励深刻影响着企业能否跨越门槛。其中，机构投资持股比反映了企业外部的资本对企业创新决策的干预；监管层总人数是内部治理架构的体现；研发人员数量代表了创新人力资本的绝对规模；而高管前三薪酬总额则揭示了核心管理层激励的集中度与公平性。这四个变量共同构成了跨越创新门槛的重要因素。

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

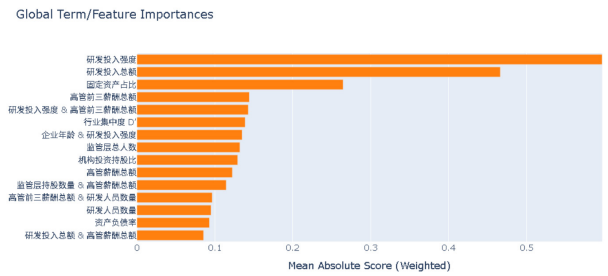


图3：特征重要性排序图

图3展示了上述四个关键变量的 EBM 形函数图，横轴为变量标准化取值，纵轴为对有“专利产出”概率的贡献。这些形状揭示了企业跨越创新门槛的具体临界条件与风险区间。

机构投资持股比的形函数图呈现出先平稳后抑制再促进的三段形态。当机构持股比例处于较低水平时，对专利产出的贡献接近于零。当持股比例上升至某个较高区间后，机构持股比对数字技术产出的影响出现了明显的负向抑制，这一发现提示，机构持股对跨越创新门槛的影响不是线性的，中等偏高的机构持股比对企业数字技术产出存在抑制现象。

监管层总人数的形函数图呈现出平稳抑制到缓慢转向促进再到达到峰值后促进效率降低的渐变形态。在监管层人数较少的阶段，对专利产出的贡献为负或接近零，说明规模过小的监管层难以形成有效的治理与决策支持；随着人数逐渐增加，贡献由负转正，且上升斜率逐步加大；当人数超过某一临界值后开始趋于平稳甚至略有下降。这表明，跨越门槛存在一个最优的监管层规模区间，过小则无力推动，过大则可能因协调成本上升而削弱决策效率。

从研发人员数量来看，形函数呈现出一段先平稳促进、达到一定数量后严重抑制的倒 V 形形态。在研发人员数量较低时，贡献微弱甚至略为负向，说明零散的研发人员难以促进企业创新发展。当研发人员数量进入中等规模后，迅速提升至强烈的正向作用，然而，一旦研发人员数量超过某个阈值，其贡献开始急剧下降甚至转为负作用，且抑制程度随人数增加而加深。这一发现说明：在跨越门槛阶段，研发人员并非越多越好。过大的研究团队可能因管理协调成本激增、研究方向分散等原因，反而降低了企业产出专利的概率。

最后关注到高管前三薪酬总额，这一变量的形函数图呈现出一条平滑的 S 形曲线。在薪酬总额较低的阶段，呈现负向影响。进入中等薪酬区间后，影响则由负转正并快速上升。当薪酬总额继续升高时，正向影响趋于平稳，不再随薪酬增加而提升，表明出现此时高管前三薪酬总额在其中的作用达到饱和。这警示企业适当的高管高薪有助于推动企业跨越专利门槛，但超出合理范围后并不会带来额外收益。



图4：单变量特征形函数图

2. 线性回归探究差异来源

探究完企业跨越门槛的因素，紧接着我们采用线性回归、随机森林与梯度提升模型，结合 SHAP 事后可解释方法，探究具有数字技术产出的企业又是通过怎样的企业结构治理来拉开创新差异。首先是线性回归的内解释作为基础比较。

表8展示了线性回归模型在基准模型与完整模型下的回归系数、z 值及显著性水平。企业结构变量中，监管层持股数量、高管薪酬总额、主营业务、政府补助、研发人员数量、监管层总人数以及高管教育背景对企业数字技术产出呈现显著的正向作用。其中监管层总人数的系数绝对值最大，为0.5151，这表明当监管层每增加一人，数字技术产出预期提升约51.5%，说明治理层人力规模在推动创新中起到了重要作用。另外监事会规模、高管前三薪酬总额、第一股东持股比、研发人员数量占比、机构投资持股比、董事会规模以及高管信息技术背景则对企业数字技术产出具有显著的负向作用。

其中存在一个矛盾的现象，即研发人员数量为正而研发人员占比为负，暗示线性模型可能无法准确刻画比例的边际效应；高管薪酬总额为正而高管前三薪酬总额为负，则预示着高管团队内部薪酬差距过大可能抑制协同创新。

表8：线性回归系数表

指标	系数	z 值	显著性	系数	z 值	显著性
控制变量						
企业年龄	0.2313	17.0497	***	0.0670	4.5785	***
总资产净利润率 ROA	0.0330	5.0950	***	0.0251	3.8732	***
总资产	−0.0900	−11.6696	***	−0.0489	−6.1944	***
企业规模	0.1443	15.6995	***	0.0114	1.0508	
营业收入增长率	0.0119	1.8655		−0.0048	−0.7710	
托宾 Q 值 D	−0.0131	−1.9368		−0.0321	−4.7481	***
资产负债率	0.0054	0.6204		−0.0233	−2.1680	*
固定资产占比	0.0922	11.2793	***	0.0586	6.8227	***
行业集中度 D’	−0.0873	−11.5097	***	−0.0914	−11.8318	***
研发投入总额	0.4243	60.1467	***	0.3258	44.3123	***
研发投入强度	0.3004	45.9443	***	0.1830	21.5187	***

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

解释变量			
监事会规模	−0.0796	−2.8659	**
监管层持股数量	0.0179	2.5111	*
高管薪酬总额	0.0655	3.1721	**
高管前三薪酬总额	−0.0507	−2.6490	**
第一股东持股比	−0.0169	−2.4302	*
主营业务	0.0816	4.9352	***
股权性质	0.0179	1.1964	
政府补助	0.0864	13.4233	***
研发人员数量	0.1304	38.2884	***
研发人员数量占比	−0.0777	−7.6554	***
监管层总人数	0.5151	9.6021	***
机构投资持股比	−0.0309	−3.6653	***
董事会规模	−0.1131	−2.5739	*
独立董事人数	0.0665	1.9006	
两职合一	0.0173	0.7234	
高管教育背景	0.1643	6.1375	***
高管信息技术背景	−0.0804	−2.9502	**

注：*、**、*** 分别代表0.05、0.01、0.001显著性水平。

（二）SHAP 事后可解释方法

SHAP 可解释机器学习（SHAPley Additive exPlanations）是一种基于合作博弈论中 SHAP 值的模型解释框架，旨在提升机器学习模型的可解释性。相较于特征重要性指标，可解释 SHAP 机器学习具有以下优势：第一，SHAP 值可以反映特征指标的贡献大小和正负向作用，为检验研究假设提供直接证据。第二，具备局部解释能力，可为局部样本生成独立的贡献值分布，从而捕捉特征在不同样本之间的异质性影响。第三，具有模型无关性，不依赖模型具体结构，可适用于各类机器学习算法。SHAP 值的计算流程如下：首先，利用加性特征归因方法将预测值 $\hat{f}(x)$ 分解为每个特征指标的贡献值之和，其中 ϕ_0 为所有样本预测均值，ϕ_i 为第 i 特征指标的 SHAP 值：

$$\hat{f}(x)=\phi_0+\sum_{i=1}^{m}\phi_i \tag{8}$$

其次，计算一个特征在所有可能的特征组合（子集）中的平均边际贡献。第 j 个因素的 SHAP 值 ϕ_i^j 计算公式如下：

$$\phi_i^j=\sum_{S\subseteq F\setminus\{i\}}\frac{|S|!(|F|-|S|-1)!}{|F|!}(\gamma_*(S\cup\{j\})-\gamma_*(S)) \tag{9}$$

其中 S 是不包含第 j 个因素的集合，共可构造出 $\sum_{S\subseteq F\setminus\{i\}}\frac{|S|!(|F|-|S|-1)!}{|F|!}$ 个不包含 j 因素的子集，$\frac{|S|!(|F|-|S|-1)!}{|F|!}$ 可以视为所有可能边际贡献的权重。$(\gamma_*(S\cup\{j\})-\gamma_*(S))_j$ 表示在变量集合 S 基础上，第 j 个因素对预测值的期望影响大小。

本案例使用到的 SHAP 可解释体现方式有以下几种：

（1）SHAP 可解释蜂群图。蜂群图能够绘制每个变量的每个样本对该模型的解释力度，从上至下按照变量总解释力度由高到低排列（即 SHAP 绝对值高低）；最右边色条表示变量数值的高低，数据点越红表示数值越高，越蓝表示数值越低；样本点分布在中轴线右边表示该样本的这个数据对结果预测是正向的影响，如果在中轴线左边则表示负向影响。

（2）SHAP 依赖图。SHAP 依赖图通过散点图的形式体现单一变量的值与 SHAP 值的关系。SHAP 蜂群图可能会有一些变量存在样本分布冗杂的情况，无法直观看出正负向影响，或者要探究其中非线性关系时，可以绘制 SHAP 依赖图。

1. 全局可解释提取关键指标

图5分别是 XGBoost 和 LighGBM 模型在 SHAP 事后可解释的蜂群图，其中研发人员数量和监管层总人数在两个模型下的排名都名列前茅，研发人数和监管层总人数较多的样本多分布在右侧，正向效应显著，说明研发人力资本以及优质的企业监管是驱动数字技术产出最核心的因素。政府补助和高管薪酬总额排名居中，政府补助在两个模型中高值样本均偏右，高管薪酬总额在 LightGBM 模型中高值样本偏右，表明外部资金支持与高管薪酬对创新产出具有强力促进作用。然而高管薪酬总额在 XGBoost 中呈现出左右均出现高值样本，低值样本集中在中间区域，两个模型呈现出不同的结果可能是受不同模型的影响，也可能是存在一定的交互作用。

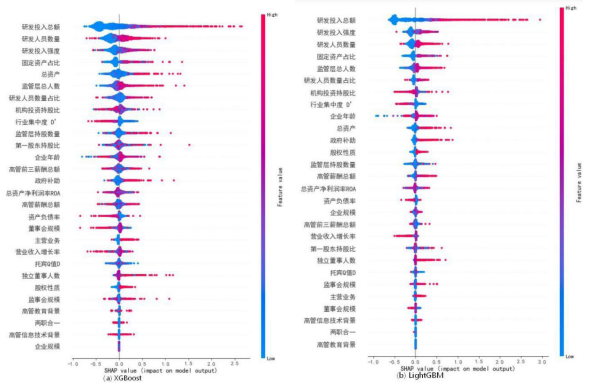


图5：SHAP 可解释蜂群图

值得特别关注的是，研发人员数量占比、机构投资持股比在整个排名中也位居前列，但这几个解释变量的 SHAP 值分布呈现明显的双向性：高值与低值样本交杂在一起，无法清晰捕捉其中正负向影响。此外，高管教育背景的影响力排名中等，但其SHAP

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

值多为正值，说明高学历管理团队普遍有益于创新。整体而言，数字技术创新的“差异鸿沟”首先源于研发人力与资金投入的差距，其次是不当的股权集中与监管层规模对创新活力的抑制。

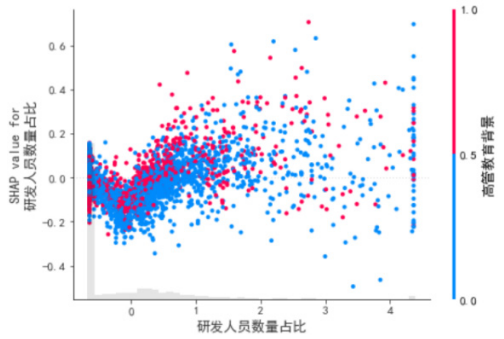
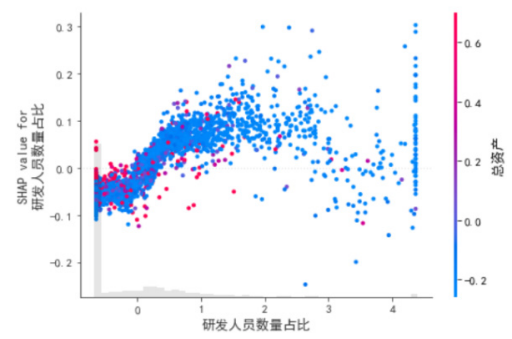
2. 单一可解释寻找非线性与交互关系

源于全局可解释下数据难以观察到非线性关系以及交互作用，于是对单个变量进行 SHAP 依赖图解释，发现以下三个指标与数字技术创新具有较为特殊的关系，且存在一定的交互项。图中数据点的颜色代表当下样本对应右侧变量的值的高低，红色点表示值较高，蓝色则相反。

（1）研发人员数量占比

表9揭示了研发人员数量占比这一变量对数字技术产出的影响，以及该变量与高管教育背景和总资产的交互作用。可以发现与 EBM 模型的形函数图存在相似的情况，即研发人员数量并非越多越好，而是存在一定峰值。并且重要的是交互作用显示较高的高管教育背景可以提升整个研发团队的水平。

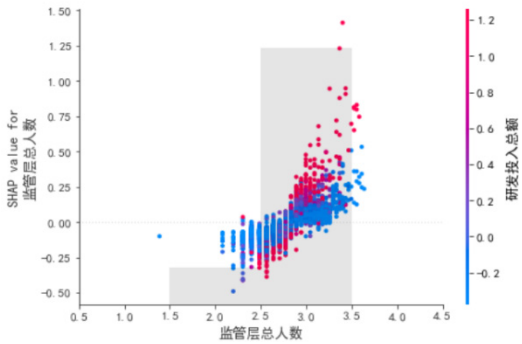
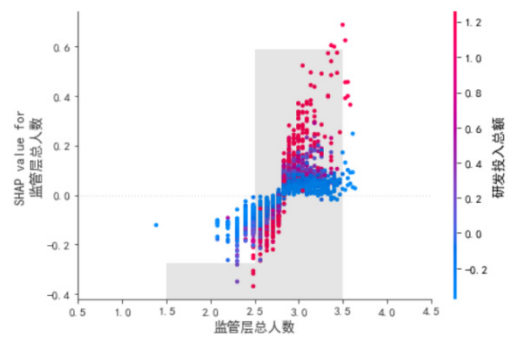
表9：研发人员数量占比分析

模型	XGBoost	LightGBM
单变量分析	随着研发人员数量占比升高，SHAP 值先是呈现‘U’式下降后升高，当占比达到一定数量时，SHAP 值开始离散并缓慢上升直至平缓。对企业数字技术创新先是抑制后促进。	与 XGBoost 模型不同的是，研发人员数量占比较少时 SHAP 值并未呈现‘U’形，而是直接上升，但后段同样变得平缓甚至出现下降，对企业技术创新产生抑制作用
交互项分析	研发人员数量占比 × 高管教育背景 高管教育背景与研发人员数量占比的交互作用下分布比较均匀，红色点的 SHAP 值分布相比蓝色点整体更高，说明高管的受教育水平较高时能够提升研发人员数量占比对企业技术创新的正向作用。	研发人员数量占比 × 总资产 总资产高值样本相比低值样本少很多，在整个依赖图在较为分散，主要集中在研发人员数量占比较少的企业，一小部分高资产企业在研发人员数量占比较少时能够降低对企业技术创新的抑制作用。
SHAP 依赖图		

（2）监管层总人数

表10呈现了监管层总人数的单一指标影响以及和研发犹如综合的交互作用。其中 XGBoost 与 LightGBM 模型的 SHAP 依赖图呈现高度相似。通过依赖图的分析可以得到以下：监管层总人数是企业数字技术创新进程中一个具有“开关”意义的治理变量，其影响力呈现指数型增长，而高研发投入总额能够放大监管层总人数的影响效应。

表10：监管层总人数分析

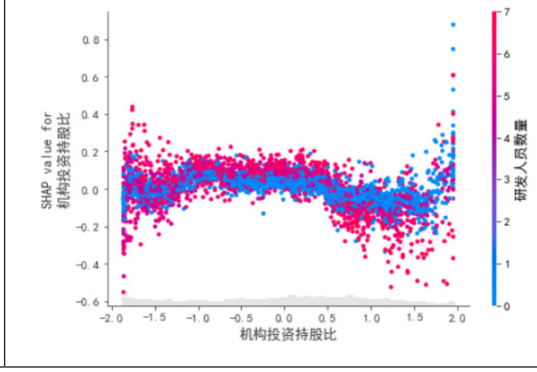
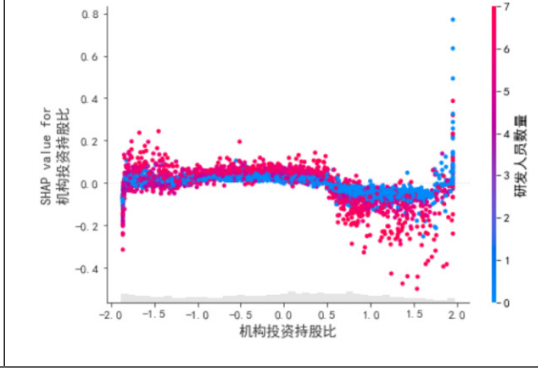
模型	XGBoost	LightGBM
单变量分析	图中显示随着监管层总人数的增加，SHAP 值呈现指数型增长，监管层总人数少的时候先是抑制数字技术创新产出，紧接着当监管层总人数增加则迅速从抑制转为促进。	
交互项分析	监管层总人数 × 研发投入总额 图中显示高研发投入在监管层影响企业技术创新过程中呈现两极分化：当监管层总人数较低，对企业数字技术创新呈现抑制效果时，高研发投入的企业数据点集中在下缘，说明此时高投入反而使得抑制效果更加明显；反之，当监管层总人数较高，对企业数字技术创新呈现促进作用是，搞研发投入的企业数据点集中在上缘，此刻高投入使得促进效果更加明显。	
SHAP 依赖图		

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

（3）机构投资持股比

表11展示的是两个模型下机构持股比的单一影响过程和其与研发人员数量的交互影响过程，且两个模型的SHAP蜂群图在该变量下也呈现出一致性。结合表内分析发现，机构持股比过高会抑制企业数字技术产出。交互作用展现初大量的研发人员数量也会放大机构持股比在其中的作用。机构持股比与研发人员数量之间存在一个“最优配合区间”，只有在机构持股适中的范围内，研发人员的规模优势才能与外部治理力量形成正向协同。

表11：机构投资持股比分析

模型	XGBoost	LightGBM
单变量分析	当机构持股比升高时，SHAP值为正，说明此时机构持股比对企业数字技术创新呈现轻微的促进作用，但当上升到一定峰值开始，SHAP值开始降到0值以下，说明在这个时候机构持股比对企业技术创新转为抑制作用。	
交互项分析	机构持股比 × 研发人员数量 研发人员数量在机构持股比对企业数字技术创新中的影响也呈现两极分化，当机构投资持股比较少或居中时，研发人员数量越多对其促进作用越有效；反之当机构持股比较高时，大量投入研发人员数量反而使得抑制效果更明显。	
SHAP依赖图		

3. 时间滚动窗口期

将时间窗口缩短至数字技术创新及企业数字化转型的发展和热门阶段，即2011–2021年期间，固定窗口期为5年，依次绘制2011–2015、2012–2016、2013–2017・・・以此类推的5年内指标特征解释性排序，形成时间滚动窗口期。

将不同周期下各解释变量的SHAP绝对值排序与完整周期下的变量SHAP绝对值排名比较，探究随着时间的推移，哪些指标始终是影响数字技术产出的关键因素，哪些指标是具有时代效应的。

综合考虑Linear、LightGBM、XGBoost模型，三者评估指标中，XGBoost的拟合优度最高，误差最小。而SHAP可解释下的交互作用LightGBM于XGBoost相差不大。因此在进行时间滚动窗口期时以XGBoost模型为代表。表12是不同时间周期内，各解释变量在所有变量中SHAP绝对值的排名。

表12：解释变量SHAP绝对值排名

变量	11–15	12–16	13–17	14–18	15–19	16–20	17–21	完整周期
第一股东持股比	4	5	5	8	8	8	11	11
股权性质	23	21	21	23	22	16	20	23
董事会规模	20	20	19	20	20	19	21	18
独立董事人数	22	24	23	21	21	23	22	22
研发人员数量	16	4	3	2	3	3	4	2
研发人员数量占比	15	16	15	13	5	9	8	7
监事会规模	21	23	22	22	23	22	23	24
高管薪酬总额	12	11	16	18	14	17	10	16
政府补助	2	2	10	12	16	18	16	14
机构投资持股比	9	14	7	6	7	6	6	8
监管层总人数	18	19	14	7	10	5	5	6
监管层持股数量	5	9	11	10	11	10	9	10
高管前三薪酬总额	8	6	13	9	13	12	12	13
两职合一	26	26	26	25	26	26	26	26
高管教育背景	24	22	27	26	27	27	27	25
主营业务	25	25	24	24	24	24	24	19
高管信息技术背景	27	27	25	27	25	25	25	27

基于时间滚动窗口与完整周期的解释变量重要性排名演变比对发现，政府补助的重要性从2011至2015年的第2位跌至后续窗口的第18位，完整周期仅列第14位，而研发人员绝对数量则从第16位跃升至第4位并长期稳居高位，这一轨迹说明政府补助和研发人员具有时代效应，并非固定效应。且企业创新动力已从政策补贴转向人力资本，表明人力结构优化正逐渐受到重视。

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

外部治理变量中，机构投资持股比稳定在第6至8位，监管层总人数从早期的第18位跃升至第6位，两者排名均保持前列或显著上升，反映出外部监督力量在创新解释体系中的地位日益强化。而传统治理变量逐渐淡出或持续影响力不足，例如第一股东持股比从早期的第4位持续下滑至第11位，股权集中的解释力明显衰退。高管薪酬总额、高管前三薪酬总额长期徘徊于第10至16位且无上升趋势。高管信息技术背景、高管教育背景、两职合一等变量重要性始终处于末端，这说明货币激励与表层治理手段并非关键解释因子，常规治理架构设置同样不是决定数字创新水平的核心。

五、研究结论

本研究使用了2000到2024年中国A股上市公司45848条年度数据，考虑到企业数字技术创新产出具有“零膨胀”和“长尾分布”同时存在的特点，把企业之间的创新差距分成了“门槛鸿沟”和“差异鸿沟”两个前后递进的层次，并搭建了一个分阶段的分析框架，先用分类模型来识别门槛，再用回归模型来解释差异。主要结论有下面几点：

第一，多种模型组合起来的方法在预测非线性很强的创新产出时，比线性模型要准确很多，而且模型评价结果在内部比较稳定；随机森林和XGBoost在预测新数据时的准确程度明显高于线性回归和Lasso，其中XGBoost在预测准确度和结果稳定性之间平衡得更好，线性模型只适合用来辅助判断系数方向；把数据分成十份反复验证后发现，各个模型评价指标在不同分法下变化很小，所以结论不会受到特定数据划分方式的影响。

第二，同一个变量在跨越门槛和扩大差异这两个阶段，发挥作用的方式有很大不同；中等规模的研发团队最容易帮助企业实现从无到有的突破，而研发人员的总人数在突破之后的差异阶段会一直起到正向作用，所以呈现出“起步阶段靠精干团队、扩大差距阶段靠人员规模”的差异特点；监管层人数规模存在明显的门槛作用，超过某个临界点之后，它的促进作用会明显增强，这一点被EBM分类模型和SHAP回归分析一起证实。

第三，用SHAP做事后解释发现了三组线性模型抓不到的非线性交互作用；监管层规模和研发投入之间存在一种“先抑制后放大”的门槛协同关系，机构持股比例和研发人员数量之间有一个最优的配合范围，研发人员占比和高管教育背景之间是正向互补的；这些交互作用说明，治理结构、股权结构和人力资本之间是否匹配，是影响数字技术创新表现的关键因素。

第四，通过时间滚动窗口分析可以看出，2011到2021年企业数字技术创新的驱动因素发生了结构性的变化；政府补助的重要性明显下降，研发人员数量和监管层规模的重要性不断上升，它们在时间上的变化轨迹像一个剪刀差，这说明创新驱动正在从依靠政策扶持转向依靠人力资本和治理能力；第一大股东持股比例和高管信息技术背景这些变量的重要性一直处于较低位置并且比较稳定，说明它们的解释能力存在结构性限制。

参考文献

[1] 钟祖昌，赵芊，余佩璇，等．企业数字化转型的技术创新效应——基于沪深A股上市公司的实证分析[J]. 科技管理研究，45(23):144-157,2025.

[2] 宁靓，王培伟．技术溢出影响企业的技术创新产出吗？——基于数字技术应用与市场竞争的联合调节效应[J]. 技术经济，44(11):110-122,2025.

[3] 白福萍，刘东慧，董凯云．数字化转型如何影响企业财务绩效——基于结构方程的多重中介效应分析[J]. 华东经济管理，36(09):75-87,2022.

[4] 丛昊，张春雨．数字技术与企业高质量创新[J]. 中南财经政法大学学报，(04):29-40,2022.

[5] 肖土盛，吴雨珊，亓文韬．数字化的翅膀能否助力企业高质量发展——来自企业创新的经验证据[J]. 经济管理，44(05):41-62,2022.

[6] 温湖炜，王圣云．数字技术应用对企业创新的影响研究[J]. 科研管理，2022,43(04):66-74.

[7] 谢婼青，方顺超，朱平芳．中小企业数字化转型的鸿沟：内在机制与弥合路径[J]. 学术月刊，58(02):71-85,2026.

[8] 孙晓华，马雪娇，孔一杰，等．中国制造业的数字鸿沟——“接入—应用—转化”的分析框架[J]. 管理世界，41(04):38-58,2025.

[9] 李柏洲，李宜函，曾经纬，等．企业如何跨越数字创新中的非连续性鸿沟？——基于博世（中国）的案例研究[J]. 科学学研究，42(06):1334-1344,2024.

[10] 徐毅，杜凤龙．数字金融、数字鸿沟与企业技术创新[J]. 哈尔滨商业大学学报（社会科学版），(06):60-71,2024.

[11] 周闯，曲植，曾思棚，等．科技金融与关键数字技术突破：来自“专精特新”企业的证据[J]. 世界经济，48(11):3-32,2025.

[12] 谢红军，齐翔．数字技术与企业价值——基于数字并购场景的研究[J]. 管理世界，41(10):146-175,2025.

[13] 陶锋，翟少轩，王峤．数字经济政策与传统企业跨界数字创新[J]. 中国工业经济，(02):118-136,2025.

[14] 沈坤荣，林剑威．链“岛”成“陆”：公共数据开放的技术创新效应研究[J]. 管理世界，41(02):83-104,2025.

[15][15] 杨鹏，尹志锋，孙宝文，等．数字技术应用与企业合作创新[J]. 经济管理，47(01):108-127,2025.

[16] 吴学福，刘振球，吴明山，等．零膨胀计数资料几种模型方法的比较研究[J]. 中国卫生统计，2020,37(3):331-334.

[17]James W M，Chandan S，Karl K，et al. Definitions，methods，and applications in interpretable machine learning.[J].Proceedings of the National Academy of Sciences of the United States of America,2019,116(44):22071-22080.

[18] 石井 一夫．5分で分かる⁉ 有名論文ナナメ読み：Marco T. Ribeiro et al.：“Why Should I Trust You?”：Explaining the Predictions of Any Classifier[J]. 情報処理，2021,62(10):568-570.

[19]Sinha K R . Book review: Christoph Molnar. 2020. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable[J].Metamorphosis，2024,23(1):92-93.

[20]Robert T . Regression Shrinkage and Selection Via the Lasso[J].Journal of the Royal Statistical Society: Series B (Methodological),2018,58(1):267-288.

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License