北京大学

硕士研究生学位论文

题目：基于 XGBoost 特征提取算法的科技型中小企业高新技术产品收入提升机制研究

姓　　名：洪玥

学　　号：2101220002

院　　系：数学科学学院

专　　业：应用统计专业

研究方向：应用统计、大数据

导　　师：董彬 教授

二〇二三 年 四 月

版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则，引起有碍作者著作权之问题，将可能承担法律责任。

摘要

创新是驱动发展的第一生产力。本文基于科技部于 2021 年收集的 7.14 万家科技型企业 2020 ~ 2021 年的数据样本，首先进行了基础的数据分析和数据挖掘，从企业的分布概况、人力结构、研发投入及产出、营业收入及产品收入等维度建立科技型企业画像。然后以原始数据集指标为基础，将指标从技术创新和企业创新管理两个维度进行划分，建立了科技型企业创新评价的指标框架，并利用原数据集进行相应的指标测算。为进一步挖掘科技型企业创新能力提升的驱动要素，以企业高新技术产品收入作为衡量企业创新能力和成长潜力的指标，先利用 XGBoost 为主的机器学习算法挖掘对高新技术产品收入预测有重要贡献的特征变量，然后进行回归建模，量化重要特征对于目标变量的内在影响机制，旨在为提升科技型中小企业创新能力提供相关政策建议，并为对科技型企业进行投融资支持提供相关参考依据。

关键词：科技型中小企业，创新能力，XGBoost，LightGBM，RandomForest, 回归分析

Research on the Promotion Mechanism of High-tech Product Revenue for Technology-oriented SMEs with a XGBoost Algorithm for Feature Importance Evaluation

Yue Hong (Applied Statistics)
Directed by: Prof. Bin Dong

ABSTRACT

Innovation is the primary driving force behind development. This article is based on a data sample collected by the Ministry of Science and Technology in 2021, consisting of 71,400 technology-based enterprises from 2020 to 2021. It first conducted basic data analysis and data mining to establish a profile of technology-based enterprises across dimensions such as distribution overview, workforce structure, research and development input and output, operating revenue, and product revenue.

Based on the original dataset, the indicators were divided into two dimensions: technological innovation and corporate innovation management. An indicator framework for evaluating innovation in technology-based enterprises was established, and corresponding indicator calculations were performed using the original dataset. In order to further explore the driving factors for enhancing the innovation capabilities of technology-based enterprises, high-tech product revenue was used as an indicator to measure innovation capabilities and growth potential. Machine learning algorithms, mainly XGBoost, were employed to identify feature variables that significantly contribute to the prediction of high-tech product revenue. Regression modeling was then conducted to quantify the intrinsic impact mechanisms of the important features on the target variable. The aim is to provide relevant policy recommendations for improving the innovation capabilities of small and medium-sized technology-based enterprises, as well as to provide reference basis for investment and financing support for technology-based enterprises.

KEY WORDS: SMEs, Innovation, XGBoost, LightGBM, RandomForest, Regression Analysis

目录

表格索引

插图索引

第一章　引言

当前，我国经济正处于从高速增长阶段转向高质量发展阶段的关键时期。在这个阶段，我们需要转变发展方式、优化经济结构并转换增长动力。实现高水平的自主发展和强国目标的关键在于创新发展。在国家创新驱动发展战略和促进大众创业、万众创新的推动下，企业作为微观主体发挥着推动科技与经济发展的重要作用。尤其是科技型企业在推动我国技术进步、科技成果转化和产业化方面发挥着突出的作用，它们是我国科技创新的核心力量，也是实现我国高水平科技自主发展和高质量发展的关键因素。因此，发挥产业政策的引导作用，推动企业创新，促进产业优化升级，利用创新驱动因素推动经济持续健康发展，具有重要的实践和政策意义。

1.1　研究背景

我国中小企业贡献了 50% 以上的税收，60% 以上的 GDP，70% 以上的技术创新，80% 以上的城镇劳动就业，90% 以上的企业数量，对于社会经济发展的重要程度不言而喻，中小企业好，我国经济才能好[1]。目前，我国正经历着重要的创新引领发展的转型期。科技型企业在这一过程中具有巨大的发展潜力，它们是创新性和突破性技术的主要孵化源，也是未来经济增长的潜在创造者。然而，当前的金融经济政策对于科技型中小企业的成长并不利。这些企业面临融资难、高死亡率等问题，这是一个不容忽视的挑战。因此，研发出一种精准支持企业创新的新型科技金融政策工具，积极识别和发现具有强大研发能力和巨大成长潜力的优质型科技企业，具有重要意义。这将有助于为企业提供精准的信用增强和信贷支持工具，引导资本、人才、技术等各类创新要素向企业聚集，切实激发微观主体的创新活力。此外，随着金融大数据和数字经济的迅猛发展，数据基础设施不断加强，数据维度不断增加，传统的统计学方法已无法满足实际应用的需求。因此，加强机器学习和人工智能方法在金融建模中的应用变得必要。

1.2　本文的主要工作

在此背景下，拟对科技部火炬中心 2021 年度收集的全国 59 家高新区的 7.14 万家科技型企业数据进行挖掘，运用机器学习方法寻找对驱动企业高新技术产品收入有重要贡献的因子，利用传统计量方法建立预测和评估模型作为辅助金融政策工具，并据此给出一定政策建议，希望对撬动社会资本、金融机构和政府资源等要素，为精准支

持具有较高创新能力和成长潜力的优质科技型企业做出贡献。

1.3 文献综述

对于科技型中小企业健康成长而言，有研究发现，技术创新能力是关键要素，与成长性呈显著正相关关系[2]，且对成长期中小企业而言，创新产出水平是对财务风险最具保护效应的因素[3]。由此可见提升创新能力，增加创新产出水平，对于科技型中小企业的健康成长具有重大意义。而影响企业创新能力和创新产出的因素众多，有些观点认为，企业研发投入程度越大，创新环境越好，专利数量越多，人力资源投入强度越大的中小企业，技术创新能力越强[2]；而对于高新技术产业，也有研究发现非研发经费的投入，比如技术改造经费等支出，同样可以对高新技术产品收入产生显著的正向影响[4]。除此之外，知识产权保护能增强企业创新动力，且私营部门企业比国有企业对知识产权保护更敏感[5]，其中有观点认为，受产业政策激励，国企组和非高新技术行业组的公司，存在为寻政策扶植在专利申请上追求“数量”而忽略“质量”的问题。[6]

除了创新能力，科技型中小企业由于其轻资产、高风险的特点，存在严重的金融排斥，资金不足成为制约科技型中小企业发展的重要因素[1]。因此挖掘驱动科技型企业的创新能力提升要素，进而预测其成长性，并为科技型企业融资构建可参照的评估体系迫在眉睫。预测科技型企业创新能力的模型本质上是经济预测类模型，某种意义上和企业的信贷评分模型的底层逻辑类似。而随着大数据的应用、信息技术的进步，与数字经济同时到来的还有高维数据任务的处理问题。面对如此丰富的数据，确定关键信息，建立一个有效且可操作的同时具有强解释力经济预测模型，并提供高质量的估计输出，特征选择是一个需要重点考虑的问题。

1.4 国内外研究现状

1.4.1 企业创新能力相关理论和指标体系建设研究现状

关于企业创新能力评价与指标体系建设的国内外相关理论，国外具有代表性的几个流派提出了不同的观点。Burgelman 和 Maidigue[7]从战略管理角度认为企业技术创新能力是企业内部一系列支持和促进技术创新战略实施的组织、技术、文化特征等的集合体。Barton[8]则从企业主体视角出发，认为企业技术创新能力由员工的知识和技能、技术系统、管理系统、科技意识和价值规范等组成。Terziovski[9]基于创新的系统集成和网络模型从创新投入、创新流程、创新产品和创新战略四个方面来测评组织的创新能力。

在国内学者方面，魏江[10]等将技术创新能力要素分解为技术创新决策能力、研发

能力、生产能力、市场营销能力和组织能力。傅家骥认为企业技术创新能力包括创新资源投入能力、创新管理能力、创新倾向、研究开发能力、制造能力和营销能力。龙艺璇[11]等人采用隐含狄利克雷分布模型（Latent Dirichlet Allocation，LDA）对 250 篇与企业创新能力评价相关的文章进行了主题调研。他们总结出目前主流文献主要围绕四个维度展开评价指标，包括技术创新投入能力、技术创新产出能力、技术创新环境支撑能力和技术创新管理能力。

1.4.2 企业技术能力评价方方面的研究现状

国内外专家学者对于企业技术创新能力评价方法的研究主要集中在层次分析法、模糊综合评判法、功效系数法、综合指数法、密切值法和指标倍数法等方面；而从方法的运用角度，可概括为单一的评价方法和相对综合的评价方法两大类。

单一评价方法是指采用单个方法对企业技术创新能力进行评价。国内外专家学者采用的方法包括以下几种：美国运筹学家 Saaty 提出的层次分析法（AHP 法）[12]，曹萍等人采用的应用网络层次分析法（ANP）[13]，Guido Capaldo 等人提出的基于模糊逻辑的方法[14]，柏昊等人采用的主成分分析法[15]，段婕[16]、吴永林[17]等人采用的因子分析法建立评价模型。

相对综合的评价方法是采用两种及以上方法相结合，取长补短、相互协调，以达到更加准确评价企业技术创新能力的目的。国内学者主要采取的综合评价方法包括以下几种：陈芝等人利用 AHP 法，并结合 BP 神经网络对技术创新能力进行评价[18]；卢怀宝等人采用二次相对评价法，形成基于 AHP 和 DEA 的综合评价法[19]；柳飞红等人采用不确定性模糊层次分析法（FAHP），然后通过简单的两两比较评判结果进行综合计算处理[20]；Wang 等人指出技术创新能力是一个复杂的、不确定性的概念，并考虑多个定量和定性标准，通过采用模糊测度和非可加模糊积分的方法，对高科技企业的技术创新能力进行评价[21]。

各种评价方法具有各自的优点和缺点，基于此本文选择相对综合的评价方法，创新性地将机器学习更高精确度的优势以及回归分析强解释性的优势相结合，用机器学习算法对全部指标进行建模预测，并作为特征变量选择的前置步骤，筛选出具有较高重要性的特征，并用线性回归进一步建模，深入重要指标对于企业高新技术产品收入的影响机制，挖掘驱动因子。

第二章　基础数据分析与指标体系构建

2.1　数据来源以及研究价值

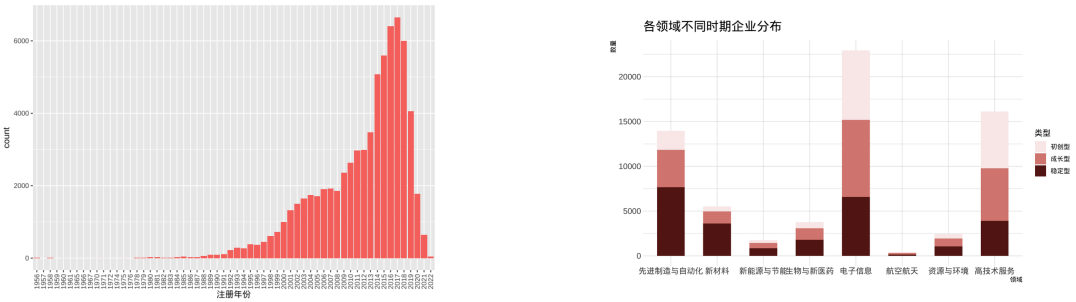
该论文数据来源于科技部。由各国家及省级高新区主要通过企业自主填报以及通过政务平台抓取两种方式对高新区内科技企业包括高新技术企业以及科技型中小企业等科技型企业进行数据采集。各单位在采集过程中遵守对数据汇聚过程的规范性与科学性要求，并对数据质量进行审核把关。最终所汇集的企业数据覆盖全国 20 个省市、59 家国家及省级高新区、八大一级技术领域共计 71494 家企业数据，经过异常值清洗，最终纳入 70013 条数据作为初始样本数据集。此数据集覆盖范围广，时效新，作为评价近两年来科技型企业发展具有极大研究价值。

2.2　数据概况以及科技型中小企业画像

原始数据集共有 71494 条数据，初步质量分析后删除了 1481 条低质量数据，剩下 70013 条作为本文分析的样本集。原数据集共涉及 21 个指标，部分指标分为 2020 年与 2021 两个年度的数据。对 21 个指标进行粗颗粒划分，主要可以分为五大类，分别是企业的属性特征、企业的人力结构、研发投入及产出、营业收入及产品收入以及其他。对数据集分别用 3-σ、和 IQR 两种方法针对其中一些存在异常值的指标进行了识别和处理之后，接下来对数据进行一些基础数据分析，初步挖掘数据集所涵盖的科技型中小企业的基本分布特征，并建立相应企业画像。

2.2.1　企业分布概况

从企业成立时间来看，大部分企业成立时间分布在 2013 到 2019 年，说明高新技术企业多处于上升成长期。从成长阶段来看，7.14 万家企业中，电子信息行业数量最多为 22924 家，占比 32.74%，其次为高技术服务行业，共 16126 家，占比 23.02%，总数排名第三的为先进制造与自动化行业，企业总数为 13962 家，占比 19.94%。从企业成长期的分布来看，7.14 万家企业中稳定型企业数量最多，占比，其次为成长型企业，占比 32.84%，最后为初创型企业，占比 37.36%。而八大领域内三大时期的企业占比结构也有较大不同，电子信息行业成长型企业占比最高 37.51%，其次为初创型企业，占比为 33.73%，最后为稳定型；先进制造与自动化领域则稳定型企业占最大比重，为 54.8%，其次为成长型企业，可以看出该行业发展较为成熟；而高技术服务行业则较为年轻化，发展势头强劲，以初创型企业数量占比最高，占比 39.43%。



(a) 企业注册年份分布情况　　(b) 企业八大领域的不同时期分布

图 2.1　企业成立时间分布概况

2.2.2　企业的人力结构

根据企业期末从业人数占比、具有研究生以上学历人员占比、科技人员占比三个维度来评价企业的人员规模和人力结构。

企业期末从业人数从 0 至 62319 人不等，平均值为 116 人，中位数为 21，前 95% 的企业的从业人数在 390 人内，整体分布属于极端的右偏态；总量企业中，25.42% 的企业从业人员期末数为 8 ~ 19 人，22.07% 的企业从业人数为 20 ~ 49 人，12.62% 的企业为 50 ~ 99 人；占比 8.10% 的 5439 家企业的从业期末人数为 100 ~ 199 人，200 人以上的企业总数为 6593，占比 9.82%。只有 1 ~ 4 人的企业有 5732 家，占比 8.54%。

具有研究生及以上学历人员从 0 至 9894 人不等，平均值为 8 人，中位数为 0，前 95% 的企业的从业人数在 20 人内，整体分布属于极端的右偏态。其中超过 50% 的企业研究生学历以上的从业人员数量为 0，研究生以上学历占比 5% 以下非 0 的企业共 10921 家，占比 15.98%；占比高于 15% 的企业共 11107 家，占比为 15.86%。

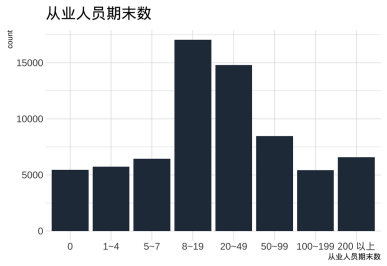
科技人员数从 0 至 15659 人不等，平均值为 27，中位数为 7，前 95% 的科技人员数在 87 人以内，整体属于极端的右偏态；从科技人员占比指标来看，总量企业研发人员占比平均值约为 30.3%，最大值约为 60%。其中研发人员占比为 0% 企业数量为 17359 家。而研发人员占比介于 40% ~ 60% 之间的企业数量为 26953 家，占比为 38.5%。

2.2.3　研发投入及产出

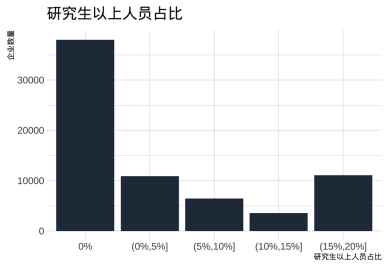
2.2.3.1　研发投入

从研发投入数据分布来看，企业的每年研发费用从 0 到 834974.2 万元不等，具有很强的头部效应。根据搜集到的企业统计，头部 1% 的企业的当年研发费用，占据了全国高新技术企业研发费用总额的 83.9%。

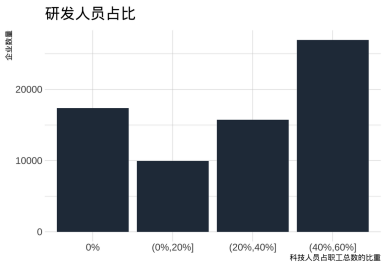
从 2021 年研发费用的描述统计来看，平均值约为 898.8 万元，中位数为 82.36 万元，最小值为 0，最大值为 83.5 亿元。前 95% 的企业研发费用在 2300 万元以内，整



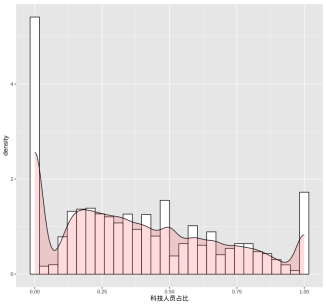
(a) 从业人员期末数分布



(b) 研究生人员占比分布



(c) 科技人员占比分布

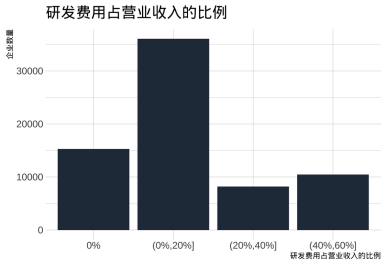


(d) 科技人员占比密度

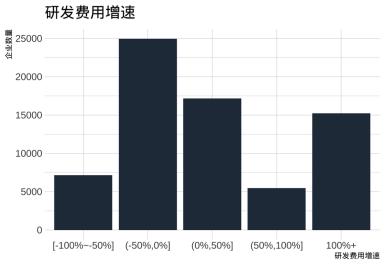
图 2.2　企业人力结构概况

体分布属于不平衡的右偏态; 从 2020 年研发费用的描述统计来看，平均值约为 1659.16 万元，中位数为 59.65 万元，最小值为 0，最大值为 1832 亿元。前 95% 的企业研发费用在 2129.9 万元以内，整体分布属于不平衡的右偏态。

从研发费用占营业收入比例来看，总量企业的研发费用占营业收入的平均比例为 15.68%，中位数为 7.17%，最小值为 0，最大值约为 60%。从研发费用增速来看，7 万家企业的研发费用增速均值为 23.055%，中位数为 6.4%, 最大值为 112.79%。负增长企业总量为 31826 家，占比 45.46%，约有 24.5% 的企业研发费用增速介于 0% ~ 50% 之间，研发费用高速增长，增长比例大于 100% 的企业一共有 15250 家，占比 21.78%。



(a) 研发费用占营业收入比例



(b) 研发费用增速分布情况

图 2.3　研发费用分布情况

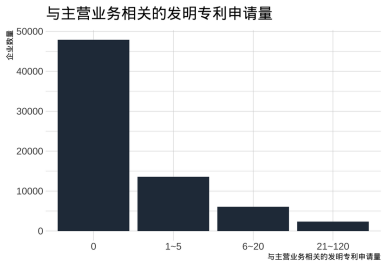
2.2.3.2 研发产出

原始数据维度中和研发产出相关的指标主要有与主营业务相关的专利申请量、与主营业务相关的 PCT 专利申请量、高新技术产品收入以及技术合同成交额。

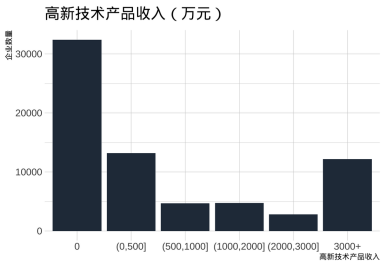
从与主营业务相关的专利申请量来看，企业每年与主营业务有关的发明专利数从 0 到 118 件不等，也具有非常明显的头部效应。根据搜集到的企业统计，头部 1% 的企业的发明专利数，占据了今年创新积分制企业发明专利总量的 84.8%。从与主营业务有关的发明专利的描述统计来看，平均值约为 3.103，中位数为 0，最小值为 0，最大值高达 118。前 95% 的企业研的发明专利申请量在 15 以内，剩余 5% 的企业在 15 到 118 不等。从专利申请数的分布来看，随着专利申请数的增加，企业数量整体呈指数下降趋势。原始数据里，有 14 家企业发明专利申请数量超过 1000。

从高新技术产品收入来看，7.14 万家企业中 46% 的企业的高新技术产品收入为 0，高新技术产品收入的平均值为 883.6 万元，前 75% 的企业的产品收入 1380 万元，企业的最大高新技术产品收入为 3496.85 万元。

从技术合同成交额来看，技术合同成交额指标为近两年企业吸纳和输出的技术合同成交总额（包含开发合作、转让合同）。两年内吸纳和输出的技术合同成交额为 0 的企业有 63776 家，占比为 91.09%；成交额为 0 ~ 100 万的有 2144 家，占比为 3.06%；成交额为 100 万至 1 亿元的一共 3723 家，占比 5.32%；技术合同成交额在 1 亿元以上的头部企业一共 370 家，占比 0.53%。



(a) 与主营业务相关专利申请量分布



(b) 高新技术产品收入分布

图 2.4 研发产出分布情况

2.2.4 营业收入及产品收入

和营业收入及产品收入相关的指标主要有 2020-2021 年度的营业收入、净利润以及所有者权益。

从 7.14 万家积分企业的营业收入分布情况来看，根据统计,2021 年高新技术企业的平均营业收入为 3233 万元，2020 年平均营业收入则为 2551.71 万元，同比增长 26.7%。

2021 年的中位数为 941.5 万元，前 75% 的数据在 4613.4 万元内，最大营业收入为 1.2228 亿元；2020 年的中位数为 660 万元，前 75% 的数据在 3685.43 万元内，最大营业收入为 9732.5 万元。相较于 2020 年，2021 的营业收入在各项统计指标上均有显著增长。从图中可看出，2020 ~ 2021 年，绝大部分企业的营业收入都小于 2500 万元。而今年的头部企业营业收入较去年整体有显著增长。

从净资产利润率来看，全量企业的平均净资产利润为 6.5% 左右，最小值约为-32.5%，最大值约为 38.5%。净资产利润率为负的企业占比为 24.1%，约 54% 的企业净资产利润率介于 0% ~ 20% 之间。

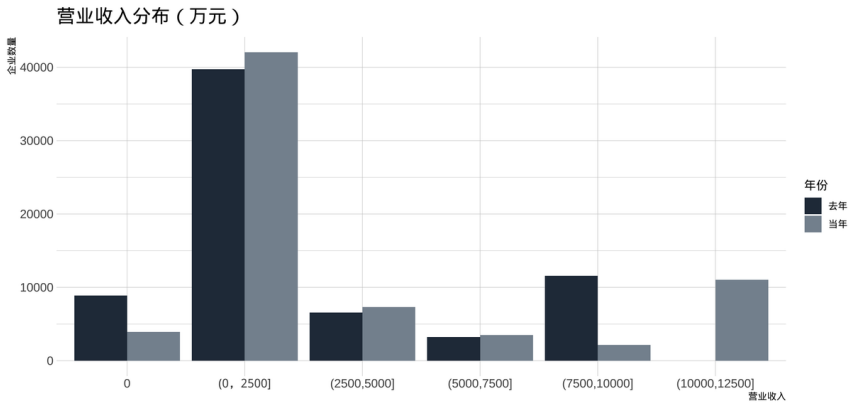
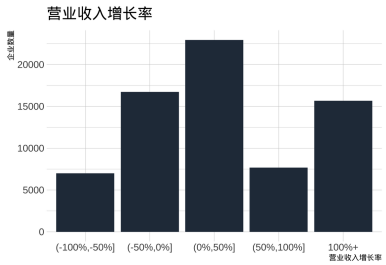
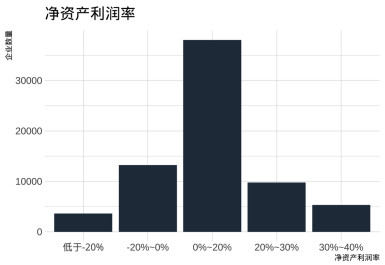


图 2.5　营业收入分布概况



(a) 营业收入增长率分布



(b) 净资产利润率分布

图 2.6　营业收入增长率及净资产利润率分布

根据期初所有者权益和期末所有者权益的箱型图可以看出，期末所有者权益各项统计指标与期初所有者权益相比有增长。期初和期末所有者权益均存在一定比例的负值，意味着企业出现资不抵债的情况。

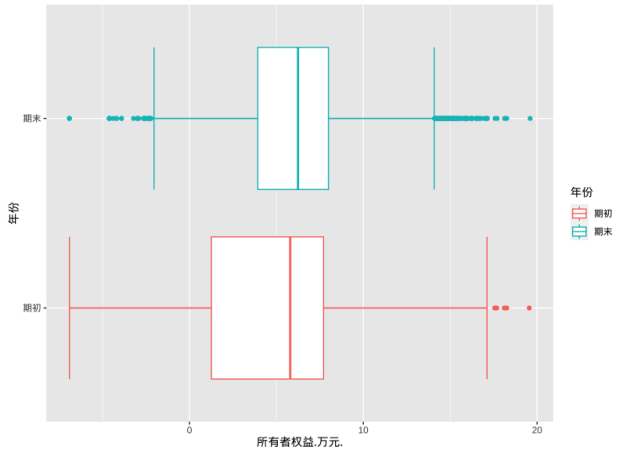


图 2.7　所有者权益分布概况

2.2.5　其他指标

其他指标还包括研发费用加计扣除所得税减免额、吸纳应届毕业生人数、承担研发和创新平台情况、承担科技计划项目情况、获得科技奖励、获得风险投资等指标。

其中，从研发费用加计扣除所得税减免额来看，减免额为 0 的企业有 49292 家企业，占比高达 70.40%，；减免额在 0 到 100 万元的企业有 14179 家，占据了总体的 20.25%；减免额在 100 万以上的企业有 6542 家，占据了总体的 9.34%，比例最小。

而企业吸纳应届毕业人数分布呈显著下降趋势，吸纳 0 人的企业高达 49981 家，占比 71.39%；吸纳 1 ~ 5 人的企业数量为 13530 家，占比 19.32%；吸纳 200 人以上的企业共计 151 家，占比 0.22%。

关于承担建设省级及以上研发或创新平台数量，2021 年最多的承担平台数量为 22，有 2016 家企业参与过研发创新平台，占比 2.88%，所有企业承担的研发和创新平台总量为 2705；2020 年最多的承担平台数量 33，有 1510 家企业参与承担了研发和创新平台，零值率 97.85%，所有企业承担的研发和创新平台总量为 2068，环比增长为 30.80%。

而关于承担省级及以上科技计划项目数量的统计数据显示，2020 年科技计划项目数量高达 33，有 955 家企业参与过承担研发和创新平台，占比 1.36%，所有企业承担的研发和创新平台总量为 1460；2021 年最多的承担平台数量 37，有 1391 家企业参与过承担研发和创新平台，占比 1.99%，环比增长 45.65%，所有企业承担的研发和创新平台总量为 2149，环比增长 47.19%。

从企业 2021 年获得科技奖励数量来看，无奖励的企业数量有 69511 家，占比 99.28%；有 502 家企业获得了科技奖励，占比 0.72% ，相较于上年增加了 36 家，增长了 7.76%。，从企业 2020 年获得科技奖励来看，无奖励的企业有 69547 家企业，占比 99.33%；有 464 家企业获得了科技奖励，占比 0.67%。

从获得风险投资金额的指标统计数据来看，只有少量企业获得了风险投资。其中无风险投资的企业有 68441 家，占比 97.75%；有风险投资的企业有 1572 家，占比 2.25%。

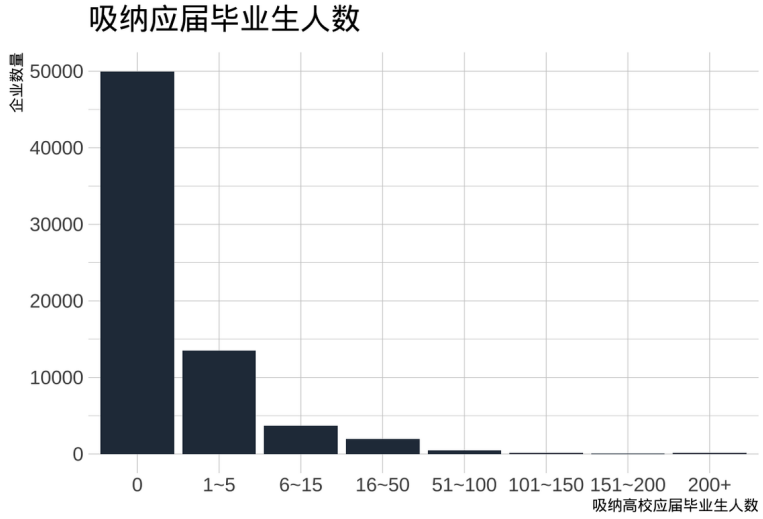


图 2.8　企业吸纳应届毕业生人数分布

2.3　科技型中小企业创新能力指标体系构建

本文对于原数据指标集的划分参照了龙艺璇等人[11]对我国科技企业创新评价相关的近几年主流文献挖掘的结果，将评价的主题主要围绕四个维度进行展开，具体包括技术创新投入能力、技术创新产出能力、技术环境支撑能力和技术创新管理能力。其中技术创新投入能力反映了企业对技术创新的关注程度，其中投入主要涉及人员、财务和物质资源，包括研发人员、研发经费和设备等，并且还包括提升人才创新素养或能力的投入。技术创新产出能力主要用于评估技术创新的成果，包括专利、论文、合同、标准以及高层次人才的拥有情况等。技术创新环境支撑能力主要考虑了企业外部环境因素，例如政府支持和区域政策，同时也包括企业内部环境，如创新文化氛围。技术创新管理能力则是企业进行技术创新的保障，主要涉及企业的创新相关政策、策略、机制和规划等方面。

基于此，以原始数据集提供的维度为基础，从技术创新维度与企业创新管理维度两个大的方面，对现有指标集进行划分，形成科技型中小企业创新能力指标体系框架，如表 3.1 所示：（1）技术创新维度。二级指标包括创新投入、创新环境支撑、创新产出共 3 项二级指标；（2）企业创新管理维度。二级指标包括企业规模、企业属性和营收能力共 3 项二级指标；

表 2.1　科技型中小企业创新能力指标体系框架

一级指标	二级指标	三级指标	变量编码	变量含义
技术创新维度	创新投入	研发费用投入强度	RD_2021 RD_2022	2020 与 2021 年度研发费用/年度营业收入
		科技人员占比	Tc_Stf_R	企业参加科技项目活动的人员总数，占企业从业人员期末人数的比例
	创新环境支撑	承担省级及以上科技计划项目数量	TP_2021 TP_2020	企业作为牵头单位近两年承担的省级及以上科技计划项目数量 (国家级科技计划项目可为项目或课题)
		承担建设省级及以上研发或创新平台数量	PF_2021 PF_2020	企业近两年获批的省级及以上重点实验室、工程中心等
		是否科小	KX	1: 是科小 0: 非科小
		是否高企	GQ	1: 是高企 0: 非高企
		研发费用加计扣除所得税减免额（万元）	Deduct	企业按照有关政策和税法规定税前加计扣除的研究开发活动费用所得税
		获得风险投资金额（万元）	RiskCap	填报期内企业获得创投机构、风险投资机构的投资额
	创新产出	与主营业务相关的发明专利申请量（件）	Ptt	填报期内企业作为第一申请人向境内外知识产权行政部门提出的与主责主业相关的发明专利申请并被受理后，按规定缴足申请费，符合进入初步审查阶段条件的件数
		与主营业务相关的 PCT 专利申请量（件）	Pct	填报期内企业作为第一申请人提出的与主责主业相关的 PCT 国际专利申请数量
		获得省级及以上科技奖励数量	TcRwd_2021 TcRwd_2020	企业作为牵头单位近两年承担的省级及以上科技计划项目数量 (国家级科技计划项目可为项目或课题)
		高新技术产品收入（万元）	Htchinc	企业生产的符合国家和省高新技术重点范围、技术领域和产品参考目录的全新型产品; 或省内首次生产的换代型产品; 或国内首次生产的改进型产品; 或属创新产品等; 具较高的技术含量和较高的附加值的产品所形成的销售收入
		企业技术合同成交额（万元）	Tcinc	企业吸纳和输出的技术合同成交总额 (开发合作、转让合同)
企业创新管理维度	企业规模	期初、期末所有者权益（万元）	eqtS eqtE	企业 2021 年度期初与期末所有者权益
		从业人员期末数	total_staff	企业 2021 年度期末从业人员总数
		具有研究生及以上学历占比	Mst_Stf_R	企业最高一级教育为研究生教育并取得毕业证书或获得硕士、博士学位证书的人员总数 (不包括肄业、结业、在读或辍学人员)，占企业从业人员期末人数的比例
		吸纳高校应届毕业生人数占比	Grad_Stf_R	2021 年度吸纳应届毕业生人数相较于当年期末从业人数占比
	企业属性	一级技术领域	F1~F9	均为 0，1 变量，F 1 ~ F9 为 1 时，分别对应高技术服务、先进制造与自动化、电子信息、资源与环境、新能源与节能、新材料、生物与新医药、航空航天、其他等 9 大领域
		所属片区	A1~A6	均为 0，1 变量，A1 ~ A6 为 1 时，分别对应华北、东北、中南、西南、西北六个地理片区
		企业成立年限	Age	截止 2022 年企业成立时长
	营收能力	营业收入（万元）	Income	企业 2021 年度营业收入
		营业收入增长率	D_income	2021 年度至 2020 年度间营业收入增长量占 2020 年度营业收入比例
		净利润（万元）	Profit	企业 2021 年度净利润
		净资产利润率	Roa_Rep	净利润/期末所有者权益
		所有者权益增长率	D_equity	期末所有者权益增长量占期初所有者权益比例

2.4　被解释变量选取

本文主要从创新产出方面来衡量企业的研发创新能力，首先考察的是能体现通过创新带来实质性收益的指标变量，即高新技术产品收入作为被解释变量，因为高新技术产品收入不仅能实质性地体现一家企业的科技创新能力，而且也能体现一家科技型企业的成长潜力，且对于企业是否能获得投融资支持具有重要的参考价值。在原数据集中，高新技术产品收入这项指标的具体含义是指企业所生产的符合国家和省高新技术重点范围、技术领域和产品参考目录的全新型产品，或是省内首次生产的换代型产品以及国内首次生产的改进型产品，亦或是具有较高技术含量和较高附加值的创新型产品所形成的销售收入。由于原始数据集中高新技术产品收入的数据分布具有较明显的右偏特征，为了提升后续进行回归分析的建模效果，故对该指标进行对数处理。

第三章　基于机器学习算法进行特征变量筛选

3.1　数据集划分以及重要特征变量筛选算法设计

在对原始数据按照表 2.1 进行三级指标构建和进一步清洗，并对异常值用 winsoring 方法进行处理后，将研究对象进一步按 8:1:1 的比例分为训练集、验证集和测试集。利用机器学习算法作为变量筛选策略，在训练集上分别利用 XGBoost、LightGBM、RandomForest 三种机器学习算法针对企业高新技术产品收入构建预测模型，并在验证集上经过网格搜索法进行参数调优。同时基于上一节所构建的科技型中小企业创新能力评价指标体系，将除高新技术产品收入之外共计 25 项三级指标下共 42 个特征，作为算法的输入特征变量，通过测试集上均方根误差 RMSE 以及判定系数 R^2 来对比评价各模型性能，最后根据最优模型所输出的特征重要程度排序，选取适当阈值，产生最终入模变量。

3.2　变量筛选模型对比

3.2.1　XGBoost 算法简介

Xgboost[5] 是一种基于梯度提升决策树 (GBDT) 的改进算法，相比 GBDT, 可以更加高效构建决策树，并在工程上实现并行运行。广泛用于解决分类和回归问题。它通过集成多个决策树模型来提高预测的准确性和性能。XGBoost 在训练过程中采用了一种称为” 梯度提升” 的技术，通过不断迭代地添加新的树模型来逐步减小预测误差。与传统的梯度提升算法相比，XGBoost 还引入了一些创新的优化策略，如自适应学习率和正则化，以提高模型的稳定性和泛化能力。此外，XGBoost 还支持并行计算，能够有效处理大规模数据集和高维特征。因此，XGBoost 算法被广泛应用于数据挖掘、预测建模和排名等各种实际问题，成为机器学习领域的重要工具之一。经过网格搜索法进行参数调优，选取模型预测正确率更高的参数组合，最终确定学习率 η=0.05，树深度 $max\_depth = 5$，$n\_estimators = 500$, 测试集 RMSE 为 6.96，$R^2 = 0.55$，变量重要性评分见图 4.1。

3.2.2　LightGBM 算法简介

LightGBM（Light Gradient Boosting Machine）是一种高效的梯度提升框架，用于解决分类和回归问题。它最初由微软研究院开发，旨在处理大规模数据集和高维特征。与传统的梯度提升算法相比，LightGBM 采用了一种称为” 基于直方图的决策树” 的策

略，以提高训练和预测的速度。LightGBM 的核心思想是将数据集划分为多个直方图，并根据直方图的信息来生成决策树。这种方式可以减少数据集的排序操作，加快了训练过程。另外，LightGBM 还引入了一些创新的优化技术，如互斥特征捆绑和直方图差分算法，以进一步提高性能。LightGBM 具有较低的内存使用和高效的并行计算能力，适用于处理大规模数据集和高维特征。它还提供了丰富的参数设置和灵活的模型调优选项，以满足不同问题的需求。由于其卓越的性能和可扩展性，LightGBM 被广泛应用于许多机器学习任务。

在经过参数调优后，最终确定 *boosting_type* 为 GBDT，使用 L2 正则项回归, 学习率 $\eta = 0.05$，树的最大深度 $\max\_depth = 5$, 迭代次数 $n\_estimators = 500$, 单棵树的叶子节点个数 $num\_leaves = 30$ ，测试集 RMSE 为 7.01，$R^2 = 0.54$。

3.2.3 随机森林算法简介

随机森林回归算法（Random Forest Regression）是随机森林（Random Forest）的重要应用分支。随机森林算法是一种集成学习方法，用于解决分类和回归问题。它结合了多个决策树模型来进行预测，被广泛应用于机器学习领域。随机森林的核心思想是通过构建多个决策树并结合它们的预测结果来做出最终的预测。每个决策树都是独立地基于随机抽样的数据集和特征子集进行训练。这种随机性的引入可以减少模型的方差，提高泛化能力。在训练过程中，随机森林通过基于特征的分割来建立决策树。它会在每个节点上从随机选择的特征子集中找到最佳分割点。这样可以增加模型的多样性，并降低决策树之间的相关性。在预测时，随机森林会将多个决策树的预测结果进行投票或平均，得到最终的预测结果。这种集成的方式可以减少模型的过拟合风险，提高预测的准确性和稳定性。随机森林算法具有许多优点，包括对异常值和噪声的鲁棒性，能够处理大规模数据集和高维特征，以及对非线性关系的良好拟合能力。它还可以提供特征重要性评估，用于特征选择和解释模型。由于其出色的性能和广泛的应用领域，随机森林算法成为了机器学习中的重要工具之一，并在实践中取得了很大的成功。

经过经过网格搜索法进行参数调优, 选取模型预测正确率更高的参数组合, 最终确定学习率 $n\_estimators = 500$，$\max\_depth = 7$, $min\_samples\_split = 6$, $\max\_depth = 7$, $min\_samples\_leaf = 5$ 测试集 RMSE 为 7.05，$R^2 = 0.54$。

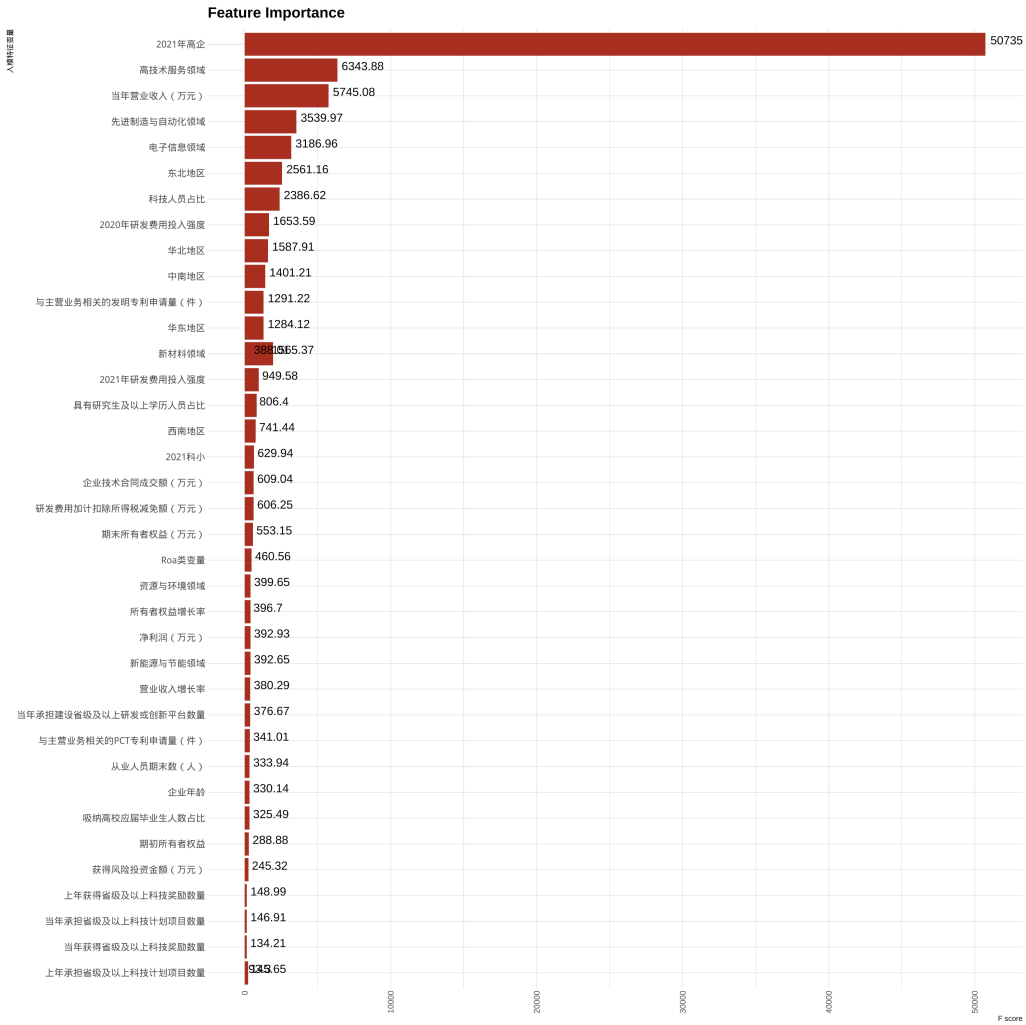


图 3.1　XGBoost 算法输出的特征重要度

3.3　机器学习算法结果比较

均方根误差（root-mean-square error）简称 RMSE，与判定系数 R^2 被作为对比三种机器学习模型预测效果的评价指标，其中 RMSE 如公式 4.1 所示，是预测值和观察值之差的二阶样本矩的平方根，常用于衡量模型预测值与观测值之间差异，该值越小，表示模型预测效果越好。而判定系数 R^2 作为回归问题的一个常用指标，通过自变量所解释的残差占总残的比例来衡量回归拟合效果，具体定义如公式 4.2，R^2 该值越大，表明模型拟合程度越好。

$$RMSE(X,h)=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h\left(x_i\right)-y_i\right)^2} \tag{3.1}$$

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \tag{3.2}$$

从三种机器学习算法结果来看，不论根据测试集上 RMSE 的表现还是回归拟合效果指标 R^2,XGBoost 模型的表现均优于 RandomForest 与 LightGBM，因此选用 XGBoost 模型的变量筛选结果作为主要参照和依据。

表 3.1　用于特征筛选的机器学习算法结果比较

	测试集 RMSE	**测试集 R^2**
算法		
XGboost	6.96	0.55
RandomForest	7.05	0.54
LightGBM	7.01	0.54

3.4　基于 XGBoost 模型的变量选择

梯度提升类算法通过构建提升树来计算各特征变量得分，从而得出每个特征对训练模型的重要性。基本思想是一个特征用于做出关键决策的次数越多，它的分数就越高。而特征变量得分的计算主要基于三个因素，分别是该过“gain”、“frequency”和“cover”[2]。其中“gain”是重要性的主要参考因素。“frequency”是所有构造树中特征出现的数量。“Cover”是特征的相对值。本文中，特征重要性由“gain”决定。

对于单个决策树，Breiman 等人[4]提出对于某个特征 X_l 的重要性得分为

$$\hat{I}_j^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 I\left(v_t = j\right) \tag{3.3}$$

其中 $J-1$ 为内部节点个数，v_t 是与节点 t 相关的特征变量，$\hat{i}_t^2$ 是节点分裂后相应的平方误差的提升。具体定义为 $i^2\left(R_l, R_r\right) = \frac{w_l w_r}{w_l + w_r}\left(\bar{y}_l - \bar{y}_r\right)^2$，其中 $\bar{y}_l$ 与 $\bar{y}_r$ 分别是节点 t 左右两棵子树的相应变量均值，w_r 与 w_l 是相应的权重和。[3]

特征的重要性取决于当该特征被随机噪声替换时预测性能是否发生显着变化。我们可以在 Xgboost 算法的训练过程中获得每个特征如何对预测性能做出贡献。显然，从图 4.1 中可看出高新技术产品收入对企业规模结构类特征这项一级指标比较敏感，该一级指标下的二级指标主要有企业属性类指标和人力结构类指标：企业属性类指标方面，企业资质是否为高新技术型企业，以及所属的一级技术领域和地理片区均具有较高特征重要性；人力结构方面，科技人员占比和高学历占比特征也具有较高重要性。此

表 3.2　最终入模变量

一级指标	二级指标	三级指标
技术创新维度	**创新投入**	2020 年研发费用投入强度 RD_2020
		2021 年研发费用投入强度 RD_2021
	创新环境支撑	科技人员占比 Tc_Stf_R
		研发费用加计扣除所得税减免额（万元） Deduct
		是否科小 KX
		是否高企 GQ
	创新产出	与主营业务相关的发明专利申请量（件） Ptt
		企业技术合同成交额（万元） Tcinc
		具有研究生及以上学历占比 Mst_Stf_R
企业创新管理维度	**营收能力**	营业收入（万元） Income
		净资产利润率 Roa_Rep

外，技术创新维度方面，创新投入、产出对于高新技术产品收入影响较大，而运营发展维度上营收能力这项二级指标对于高新技术产品收入的贡献较大。

最后根据 XGBoost 预测模型的特征重要度结果，共入选了包含研发费用投入强度、与主营业务相关的发明专利申请量、企业技术合同成交额、科技人员占比、研究生以上学历占比、营业收入、净资产利润率、研发费用加计扣除所得税减免额以及高企、科小共计 11 个三级指标作为最终入模特征变量，进一步建模探究以上特征对于高新技术产品收入影响的内在机制。

第四章　创新能力评价模型设计与实证分析

4.1　解释变量

根据 XGBoost 预测模型的特征重要度结果，共入选了包含研发费用投入强度、与主营业务相关的发明专利申请量、企业技术合同成交额、科技人员占比、研究生以上学历占比、营业收入、净资产利润率、研发费用加计扣除所得税减免额以及高企、科小共计 11 个三级指标入模，进一步建模探究以上特征对于高新技术产品收入影响的内在机制。首先对除高小和科企之外的 9 个变量做进一步的数据挖掘。

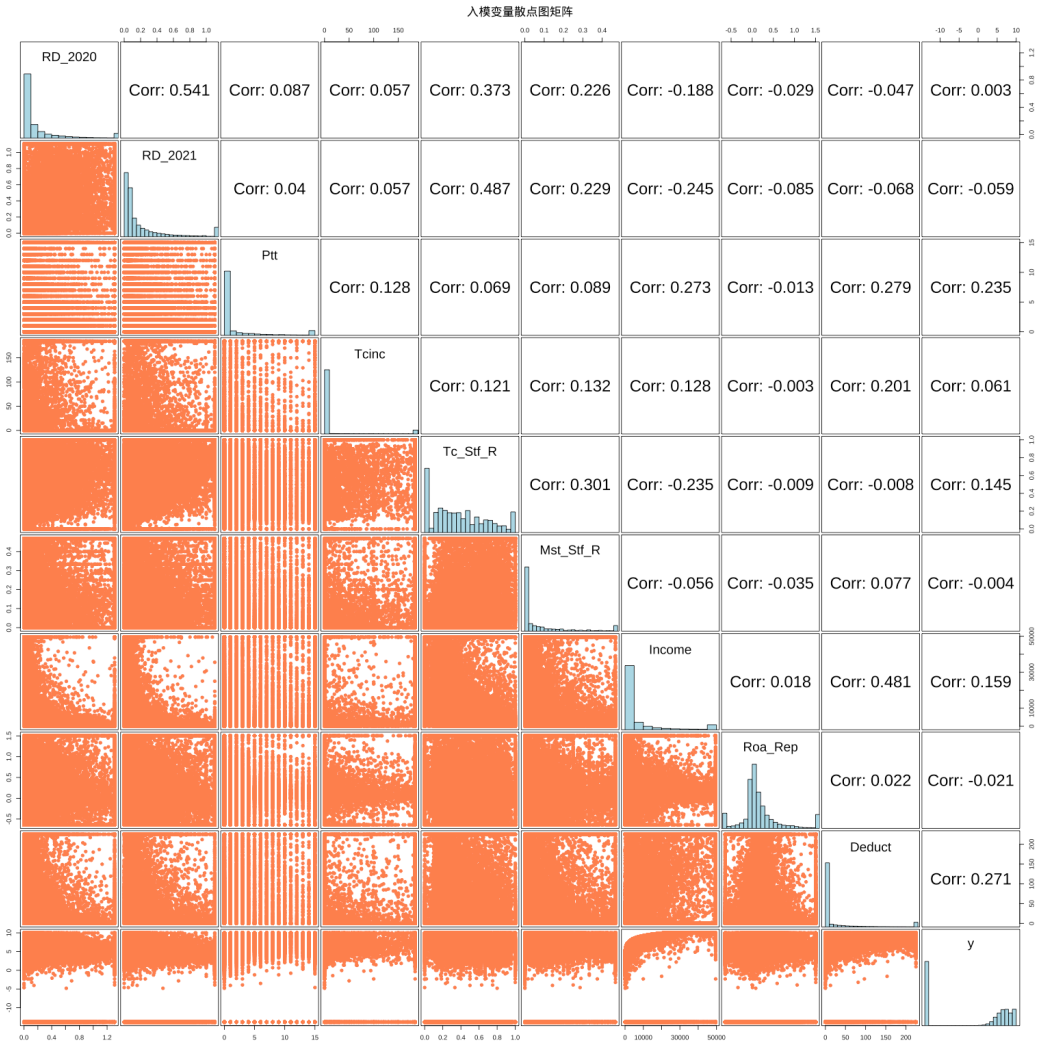


图 4.1　变量之间的相关性分析

根据图 5.1，首先可从上三角的相关系数矩阵以及下三角的散点图矩阵看出所有变

量之间的相关性系数均小于 0.7，变量两两之间不存在较强相关性。其次图 5.1 的对角线上个变量的分布直方图，可以看到非高企科小其余 9 个解释变量均存在较高的 0 值率，以及右偏的分布特征，而被解释变量高新技术产品收入的数值同样存在较多的 0 值，以及呈现双峰的分布特征，右峰呈左偏态。

表 5.1 呈现了入模解释变量的描述性统计结果。

表 4.1　解释变量描述性统计

变量	平均值	标准差	最小值	25% 分位数	50% 分位数	75% 分位数	最大值
2020 年研发费用投入强度 RD_2020	0.192	0.32	0.0	0.0	0.063	0.195	1.308
2021 年研发费用投入强度 RD_2021	0.196	0.282	0.0	0.035	0.076	0.224	1.104
与主营业务相关的发明专利申请量（件） Ptt	1.943	4.041	0.0	0.0	0.0	2.0	15.0
企业技术合同成交额（万元） Tcinc	12.576	44.096	0.0	0.0	0.0	0.0	184.188
科技人员占比 Tc_Stf_R	0.38	0.301	0.0	0.139	0.333	0.6	1.0
具有研究生及以上学历占比 Mst_Stf_R	0.074	0.128	0.0	0.0	0.0	0.089	0.469
是否科企 KX	0.536	0.499	0.0	0.0	1.0	1.0	1.0
是否高小 GQ	0.703	0.457	0.0	0.0	1.0	1.0	1.0
营业收入（万元） Income	6623.743	12773.832	0.0	228.818	1150.0	5199.048	49529.84
净资产利润率 Roa_Rep	0.169	0.45	-0.639	0.0	0.068	0.26	1.505
研发费用加计扣除所得税减免额（万元） Deduct	25.495	59.721	0.0	0.0	0.0	8.16	225.08

4.2　控制变量

除上述 11 个变量外，根据 XGBoost 算法特征重要度的排序结果，一级技术领域和地理片区也对预测具有显著影响。对一级技术领域、地理片区变量和目标预测变量即高新技术产品收入对数值进行交叉分析，可视化结果如图 5.2、5.3 所示。可看到不同技术领域的高新技术产品收入对数值分布具有较大差异，尤其是高技术服务、先进制造与自动化、电子信息以及新材料领域和其他四大领域之间的分布差异较大。同时不同地理片区上高新技术产品收入分布也存在差异，中南、西南、华东地区的高新技术收入显著高于西北、华北和东北地区。

故在实际模型中将二者进行控制。一级技术领域特征引入 F1 ~ F9 虚拟变量，分别对应高技术服务、先进制造与自动化、新材料、新能源与节能、生物与新医药、电子信息、航空航天、资源与环境与其他共九大类一级技术领域；地理分区上则引入 A1 ~ A6 分对应华东、华北、中南、西北、西南、东北六大地理分区。

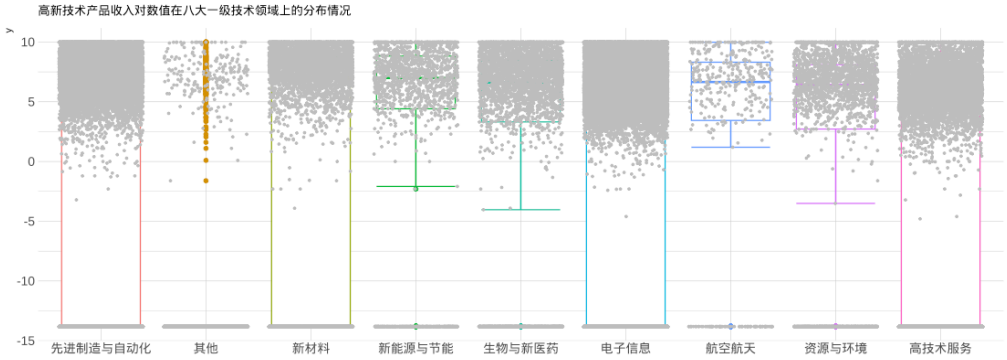


图 4.2　八大技术领域上高新技术产品收入的对数值分布

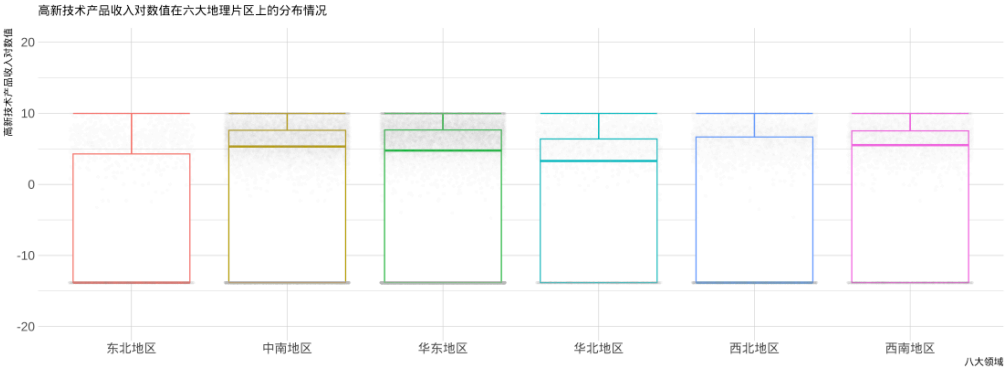


图 4.3　八大技术领域上高新技术产品收入的对数值分布

4.3　回归模型设定

为进一步探究并量化上述具有较高重要度的 11 个特征对于高新技术产品收入的影响机制，将高新技术产品收入的对数值作为被解释变量，一级技术领域和地理分区作为控制变量，建立回归模型，模型设计如（5.1）所示：

$$\begin{aligned} y =& \beta_0 + \beta_1 * RD\_2020 + \beta_2 * RD\_2021 + \beta_3 * Ptt + \beta_4 * Tcinc \\ &+ \beta_5 * Tc\_Stf\_R + \beta_6 * Mst\_Stf\_R + \beta_7 * Income + \beta_8 * Roa\_Rep \\ &+ \beta_9 * Deduct + \sum_{m=1}^{9} \beta_m F_m + \sum_{n=1}^{6} \beta_n A_n + \epsilon \end{aligned} \tag{4.1}$$

4.4 模型结果

得到的回归结果如表 5.2 所示，其中模型（1）同时控制了一级技术领域和地理分区的双向固定效应模型，其 R^2 为 0.398，测试集上 RMSE 为 8.00；模型（2）是仅控制了地理分区的固定效应模型，R^2 为 0.372，测试集上 RMSE 为 8.19；模型（3）是仅控制一级技术领域的固定效应模型，其 R^2 为 0.392，测试集上 RMSE 为 8.04。从判定系数 R^2 和测试集上的均方根误差 RMSE 的表现综合来看，模型（1）的表现最优。

表 4.2　对高新技术产品收入的回归结果

变量	**模型（1）**	**模型（2）**	**模型（3）**
截距项	-11.4989*** （0.000）	-12.3883*** （0.000）	-10.3733*** （0.000）
2021 年研发费用投入强度	-3.9384*** （0.000）	-4.5424*** （0.000）	-3.9091*** (0.000)
2020 年研发费用投入强度	0.1156 （0.39）	-0.1248 （0.362）	0.2180 (0.106)
与主营业务相关的发明专利申请量（件）	0.1806*** （0.000）	0.2091*** （0.000）	0.2089*** (0.000)
企业技术合同成交额	-0.0039*** （0.000）	-0.0064*** （0.000）	-0.0047*** (0.000)
科技人员占比	1.8150*** （0.000）	0.6002*** （0.000）	1.5700*** (0.000)
具有研究生及以上学历占比	0.66** (0.03)	0.0683 （0.825）	0.9884*** (0.001)
科小	2.335*** (0.000)	2.3158*** （0.000）	2.3934*** (0.000)
高企	10.1350*** (0.000)	10.8966*** （0.000）	10.2268*** (0.000)
营业收入（万元）	0.00007*** (0.000)	7.615e-05 （0.000）	6.978e-05 (0.000)
净资产利润率	-0.1204 (0.128)	-0.1435 (0.075)	-0.1624** (0.041)
研发费用加计扣除所得税减免额（万元）	0.0192*** (0.000)	0.0188*** (0.000)	0.0195*** (0.000)
固定效应（一级技术领域） 固定效应（地理分区）	控制 控制	- 控制	控制 -
R^2	0.398	0.372	0.392
测试集 RMSE	8.00	8.19	8.04

4.5　模型结果分析

4.5.1　2021 年研发投入强度

从一级指标技术创新维度来看，创新投入项二级指标下只有 2021 年的研发费用投入强度变量的系数在 1% 置信水平下显著为负，模型（1）中 $\beta_2 = -3.9384$, 说明 2021 年度的研发投入强度对于高新技术产品收入是一个负向因子。即 2021 年度研发投入强度每增加一个单位，同期高新技术产品收入的对数值就会下降约 3.9 个单位。

4.5.2　与主营业务相关的发明专利申请量

与主营业务相关的发明专利申请量的回归系数为 0.1806，该变量对于高新技术产品收入是促进因子。即主营业务相关的发明专利申请量增加 1 个单位，企业高新技术产品收入的对数值就会增加 0.1806 个单位。主营业务相关专利申请量为技术创新维度中的创新产出指标，表明科技型企业若积极进行创新产活动出，随着专利申请量的增加，不仅有助于推动企业技术进步，提升竞争优势，还可获得实质性高新技术产品收入方面的增益。

4.5.3　企业技术合同成交额

企业技术合同成交额的回归系数为-0.0039，且在 1% 的置信水平之下显著为负，该变量对于企业的高新技术产品收入而言是负向因子。即企业技术合同成交额每增加一个单位，企业的高新技术产品收入对数值就会减少 0.0039 个单位。企业技术合同成交额为技术创新维度下的创新收益指标，其含义为企业吸纳和输出的技术合同成交总额，涵盖开发合作，转让合同等形式，是企业进行与科技产品相关的商业活动的度量指标。应注意本次高新技术合同成交额数据与高新技术产品收入数据的采集时间，二者均为企业 2021 年数据，由于在实际经济活动中，技术合同成交额属于企业的应收账款，存在未到账以及一定程度的不确定因素，对高新技术产品收入的提升存在时间上的滞后效应。回归结果系数表明，在科技产品上的开发和转让的活跃度并不一定能带来同时期正向的高新技术产品收入。

4.5.4　科技人员占比

科技人员占比回归系数为 1.8150，在 1% 的置信水平之下显著为正，改变量对于企业的高新技术产品收入而言是促进因子。即企业科技人员占比每增加一个单位，企业的高新技术产品收入对数值将会增加 1.815 个单位。科技人员占比为企业规模结构特征中的人力结构下的指标，该回归结果表明企业在人力结构配置中若积极增加实际参加科技活动的人员数，将有助于提升高新技术产品带来的实际收入。

4.5.5 具有研究生及以上学历人员占比

具有研究生及以上人学历人员占比的回归系数为 0.66，在 5% 的置信水平之下显著为正，改变量对于企业的高新技术产品收入而言是促进因子。该指标依然为企业规模结构特征中的人力结构下的指标，回归结果表明，通过提升企业中研究生学历人员占比，优化人力结构，将正向促进于能带来实际竞争力的高新技术产品收入的增加。

4.5.6 科小与高企

科小的回归系数为 2.335，同时高企的回归系数为 10.1350，二者在 1% 的置信水平之下均显著为正。科小和高企两个变量均为企业规模结构指标下的企业属性指标，这两项指标不仅在回归模型中对高新技术产品收入而言为显著的促进因子，在 XGBoost 模型中同样具有较高的特征重要度。科小的全称为科技型中小企业，是科技部、财政部、国家税务总局于 2017 年为推动大众创业万众创新，加大对科技型中小企业的精准支持力度而推出的针对中小型科技型企业的资格认证。高企的全称为高新技术企业，是科技部、财政部、国家税务总局于 2016 年推出的资格认证，获得这两项资质均有助于享受税收优惠政策等一系列支持服务。回归结果表明企业具有这两项资质，尤其是高新技术企业的资质认定，将极大地促进高新技术产品收入的提升，对科技型企业的扶持策略有效地为科技型企业带来了创新收益。

4.5.7 营业收入

营业收入的回归系数为 0.0007，在 1% 的置信水平之下均显著为正。说明企业的营业收入水平越高，高新技术产品收入也会随之提升。因此提升企业的营收能力，也将正向作用于高新技术产品的营收。

4.5.8 研发费用加计扣除所得税减免额

研发费用加计扣除所得税减免额的回归系数为 0.0192，在 1% 的置信水平之下均显著为正，则该因子为正向促进因子，即研发费用加计扣除所得税每增加一个单位，高新技术产品收入的对数值将会提升 0.0192 个单位。表明出台有利于科技型企业的税收优惠政策将有助于提升科技型企业的高新技术产品收入。

第五章　总结和展望

5.1　结论及政策建议

本文选取科技部2021年度收集的7.14万家科技型企业数据作为样本，从技术创新维度与企业创新管理维度对原数据集指标进行划分，形成了科技型中小企业创新能力指标体系框架，并以高新技术产品收入作为企业创新能力和成长潜力的评价指标，利用XGBoost模型挖掘出了11个具有较高重要度的特征因子，并进行回归建模，进一步量化探究各项因子对高新技术产品收入的影响机制，得出结论是，对高新技术产品收入具有提升效应的是与主营业务相关的发明专利申请、科技人员和具有研究生以上学历人员占比、具备高新技术企业与科技中小型企业认证资质、营业收入以及研发费用加计扣除所得税减免额。

从以上结果可以看出政府出台的针对科技型企业的扶植政策和认证机制可以显著激励科技型中小企业的创新能力，带来高新技术产品收入的正向增长；同时企业的营收能力和高素质的优良人才配置也是增强创新能力的关键要素，因此，吸引高学历人才以及科技人员等专业人才加入的优化人力结构措施是提升企业创新能力的关键之一；积极进行创新产出，提升与企业主营业务相关的专利数量，也是增加高科技产品收入的直接手段。综上，为提高科技型中小企业的创新能力，政府应积极出台相应利好政策，而科技型中小企业自身则应该积极优化人员结构，提高营收能力，增加核心专利产出数量。

5.2　本文局限性及未来进一步工作

对于结果的评价上，本次模型结果显示企业同期的研发投入强度和企业技术合同成交额对于同时期的高新技术产品收入的效应是负面的，经验表明研发投入和企业技术合同成交额应该是正向因素，对该现象的合理推测和猜想为这二者对于高新技术产品收入的提升存在一定滞后效应，需要更长的时间序列数据进行验证。未来希望能进一步收集数据对此进行深入检验。

除此之外，处于不同生命周期的科技型企业因内部财务状况和外界资源等诸多因素限制，可能具有不同的创新能力驱动要素，目前本文的工作未对所研究的科技型企业进行相应的生命周期划分，会在未来的工作中进一步细化，建立针对初创期、成长期、稳定期科技型中小企业相对应的评价模型，并探究处于处于不同生命周期的企业的创新驱动要素特点差异，进一步精细化现有工具。此外，本文对于模型特征选择的

算法比较仅局限于机器学习算法自身，而未与传统变量选择方法进行比较，将来会进一步细化该工作。

参考文献

[1] 徐海龙, 王宏伟. 科技型中小企业全生命周期金融支持研究——基于风险特征的分析视角[J]. 科学管理研究, 2018, 36(3): 56-59.

[2] 陈晓红, 马鸿烈. 中小企业技术创新对成长性影响——科技型企业不同于非科技型企业?[J]. 科学学研究, 2012, 30(11): 1749-1760.

[3] 闵剑, 李佳颖. 生命周期视角下中小企业财务风险评估研究——基于生存分析模型[J]. 财会通讯, 2021(02): 146-150.

[4] 谢子远, 黄文军. 非研发创新支出对高技术产业创新绩效的影响研究[J]. 科研管理, 2015, 36(10): 1-10.

[5] FANG L H, LERNER J, WU C. Intellectual Property Rights Protection, Ownership, and Innovation: Evidence from China[J]. The Review of Financial Studies, 2017, 30(7): 2446-2477.

[6] 黎文靖, 郑曼妮. 实质性创新还是策略性创新?——宏观产业政策对微观企业创新的影响[J]. 经济研究, 2016, 51(4): 60-73.

[7] BurgelmanR.，MaidiqueM.A.，WheelwrightS.C. Strategie Management of Technology and Innovaiton[J]. NewYork:Mc—Graw-Hil, 2004.

[8] Barton. Corecapabilities and corerigidities: aparadoxinman.agingnew productdevelopment[J]. Stategic Management Jourhal, 1992.

[9] M. T. Achieving performance excellence through an integrated strategy of radical innovation and continuous improvement[J]. Measuring Business Excellence, 2002(6): 5-14.

[10] 魏江, 郭斌, 许庆瑞. 企业技术能力与技术创新能力的评价指标体系[J]. 中国高新技术企业评价, 1995(5): 33-38.

[11] 龙艺璇, 高钰涵, 翟夏普, 等. 我国企业技术创新能力评价指标体系研究现状分析——基于文献计量与主题模型[J/OL]. 科学观察, 2023, 18(2), 24: 24-31. https://manu56.magtech.com.cn/kxgc/CN/10.15978/j.cnki.1673-5668.202302005. DOI: 10.15978/j.cnki.1673-5668.202302005.

[12] SAATY R W. The Analytic Hierarchy Process—What It Is and How It Is Used[J]. Mathematical Modelling, 1987, 9(3): 161-176 [2022-08-18]. DOI: 10.1016/0270-0255(87)90473-8.

[13] 曹萍, 陈福集. 基于 ANP 理论的企业技术创新能力评价模型[J]. 科学学与科学技术管理, 2010(2): 67-71.

[14] CAPALDO G, IANDOLI L, RAFFA M, et al. The evaluation of innovation capabilities in small software firms: A methodological approach[J]. Small Business Economics, 2003, 21: 343-354.

[15] 柏昊, 杨善林, 钟金宏. 基于主成分分析法的制造业产业技术创新评价模型及应用[J]. 合肥工业大学学报: 自然科学版, 2007, 30(3): 322-325.

[16] 段婕, 刘勇. 基于因子分析的我国装备制造业技术创新能力评价研究[J]. 科技进步与对策, 2011, 28(20): 122-126.

[17] 吴永林, 赵佳菲. 北京高技术企业技术创新能力评价分析[J]. 企业经济, 2011(3): 21-23.

[18] 陈芝, 张东亮, 单汨源. 基于 BP 神经网络的中小企业技术创新能力评价研究[J]. 科技管理研究,

2010(2): 56-58.

[19] 卢怀宝, 冯英浚, 曲世友, 等. 企业技术创新能力的二次相对评价法[J]. 大庆石油学院学报, 2002, 26(1): 90-93.

[20] 柳飞红, 傅利平. 基于 FAHP 的企业技术创新能力评价指标权重的确定[J]. 统计与信息论坛, 2009, 24(2): 24-28.

[21] WANG C H, LU I Y, CHEN C B. Evaluating firm technological innovation capability under uncertainty[J]. Technovation, 2008, 28(6): 349-363.

[22] 林毅夫. 发展战略、自生能力和经济收敛[J]. 经济学（季刊), 2002, 1(2): 269-300.

[23] 唐雯, 陈爱祖, 饶倩. 以科技金融创新破解科技型中小企业融资困境[J]. 科技管理研究, 2011, 31(7): 1-5.

[24] WEST D. Neural network credit scoring models[J]. Computers & Operations Research, 2000, 27(11): 1131-1152.

[25] JENSEN H L. Using Neural Networks for Credit Scoring[J]. Managerial Finance, 1992, 18(6): 15-26. DOI: 10.1108/eb013696.

[26] YANG Y, PANG Y, HUANG G, et al. The Knowledge Graph for Macroeconomic Analysis with Alternative Big Data[Z]. 2020. arXiv: 2010.05172[cs,econ,q-fin]. DOI: 10.48550/arXiv.2010.05172.

[27] BIAU G, SCORNET E. A random forest guided tour[J]. TEST, 2016, 25(2): 197-227. DOI: 10.1007/s11749-016-0481-7.

[28] BOULESTEIX A L, JANITZA S, KRUPPA J, et al. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics[J]. WIREs Data Mining and Knowledge Discovery, 2012, 2(6): 493-507. DOI: 10.1002/widm.1072.

[29] PROBST P, WRIGHT M N, BOULESTEIX A L. Hyperparameters and tuning strategies for random forest[J]. WIREs Data Mining and Knowledge Discovery, 2019, 9(3): e1301. DOI: 10.1002/widm.1301.

[30] BELGIU M, DRĂGUŢ L. Random forest in remote sensing: A review of applications and future directions[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 114: 24-31. DOI: 10.1016/j.isprsjprs.2016.01.011.

[31] DAOUD E A. Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset[J]. International Journal of Computer and Information Engineering, 2019, 13(1): 6-10.

[32] KE G, MENG Q, FINLEY T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree [C]//Advances in Neural Information Processing Systems: vol. 30. Curran Associates, Inc., 2017.

[33] SUN X, LIU M, SIMA Z. A novel cryptocurrency price trend forecasting model based on LightGBM [J]. Finance Research Letters, 2020, 32: 101084. DOI: 10.1016/j.frl.2018.12.032.

[34] CHEN T, GUESTRIN C. XGBoost: A Scalable Tree Boosting System[C]//KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2016: 785-794. DOI: 10.1145/2939672.2939785.

[35] FRIEDMAN J H. Stochastic gradient boosting[J]. Computational Statistics & Data Analysis. Nonlinear Methods and Data Mining 2002, 38(4): 367-378. DOI: 10.1016/S0167-9473(01)00065-2.

[36] HASTIE T. The elements of statistical learning : data mining, inference, and prediction[M]. 2001.

[37] LOH W Y. Classification and regression trees[J]. WIREs Data Mining and Knowledge Discovery, 2011, 1(1): 14-23 [2023-04-20]. DOI: 10.1002/widm.8.

[38] FRIEDMAN J H. Greedy Function Approximation: A Gradient Boosting Machine[J]. The Annals of Statistics, 2001, 29(5): 1189-1232.

[39] ZHENG H, YUAN J, CHEN L. Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation[J]. Energies, 2017, 10(8): 1168. DOI: 10.3390/en10081168.

[40] 马倩倩，孙东旭，石金铭，何贤英，翟运开, YUNKAI M Q D J X. 基于支持向量机与 XGboost 的成年人群肿瘤患病风险预测研究[J]. 中国全科医学, 2020, 23(12): 1486. DOI: 10.12114/j.issn.1007-9572.2020.00.066.

[41] WANG Y, NI X S. A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization[Z]. 2019. arXiv: 1901.08433[cs,stat]. DOI: 10.48550/arXiv.1901.08433.

[42] BEN JABEUR S, STEF N, CARMONA P. Bankruptcy Prediction using the XGBoost Algorithm and Variable Importance Feature Engineering[J/OL]. Computational Economics, 2023, 61(2): 715-741(2023-02-01) [2023-04-20]. https://doi.org/10.1007/s10614-021-10227-1. DOI: 10.1007/s10614-021-10227-1.

[43] SHI X, WONG Y D, LI M Z F, et al. A feature learning approach based on XGBoost for driving assessment and risk prediction[J]. Accident Analysis & Prevention, 2019, 129: 170-179 [2023-04-20]. DOI: 10.1016/j.aap.2019.05.005.

[44] 段木林. 企业技术创新能力评价的综述研究[J]. 管理观察, 2013(15): 22-22.

[45] 杨云, 蔡德军. 企业技术创新能力评价的指标与方法综述[J]. 铜陵学院学报, 2012, 11(2): 41-44.

[46] LOWEL S. W.EvlauatingtheTechnical Operation[J]. Stategic Management Jourhal, 1988.

致谢

在北大学习生活了短短两年，感谢这期间遇见的给予过我指导或在课堂上激起我求知欲的老师以及优秀热诚的同学。两年间不论是学识、专业技能还是心智，都在各种历练考验中提升了很多，感谢这个园子给我带来的难忘的记忆，也感谢爱我的人和我爱的人在这期间所有的支持与陪伴。

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：　　日期：　　年　　月　　日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：　　导师签名：

日期：　　年　　月　　日