Python 数据分析课程设计报告

题　　目: 专精特新中小企业科创水平影响因素分析

院　　系: 人工智能学院

专　　业: 数据科学与大数据技术

年级班别: 21 级数据科学与大数据技术（专升本）1 班

学　　号: 221470101

学生姓名： 包旻宇

授课教师: 陶建敏

提交日期： 2024 年 11 月 24 日

摘要

本研究针对“专精特新”中小企业，通过数据挖掘与建模预测，探索影响企业科创水平的关键因素，并为企业的创新能力提供科学量化的评估方法。研究基于企业的基础信息与知识产权相关数据，进行了全面的数据预处理与特征工程，筛选出与企业评分密切相关的特征变量。采用梯度提升树（Gradient Boosting Decision Tree, GBDT）作为核心模型，对企业的评分进行回归预测，模型在多个误差指标（MSE、RMSE、MAE）和决定系数（R^2）上均表现优异，预测误差较小，稳健性强，显示出该模型在处理中小企业科创水平评分任务中的高度适配性。

在数据可视化方面，研究运用 Python 的可视化库（如 Plotly、Seaborn 等）以及 FineBI 工具，对企业数据进行了多维度统计分析和直观展示。具体分析了基础信息、知识产权指标与企业科创评分之间的关联规律，揭示了如有效发明专利数、软件著作权数等特征对评分的显著影响。通过分析区域分布、特征重要性以及评分预测的相对误差与残差分布，进一步验证了数据与模型的科学性与解释力，为政策制定者、企业管理者以及行业研究者提供了重要的决策依据。

为提升研究成果的实际应用价值，研究基于 Streamlit 框架设计了交互式 Web 界面，实现了企业评分的便捷预测功能。用户只需输入如知识产权数量、企业类型等相关信息，即可获得模型预测的评分结果。该界面不仅提供了便捷的预测服务，还直观展示了企业评分的特征重要性分析，为企业了解自身优势与改进方向提供了实用参考。

本研究的创新点在于结合企业基础信息与知识产权指标，构建了适用于“专精特新”中小企业的评分预测模型，并通过 Web 应用实现了模型成果的实际转化。研究结果不仅为企业创新能力的精准评估提供了技术支持，还为未来的行业趋势分析、政策优化以及企业创新路径规划提供了新的思路。

关键词：专精特新中小企业、数据挖掘、知识产权、梯度提升树

Abstrac

This study focuses on "specialized, refined, unique, and innovative" small and medium-sized enterprises (SMEs), using data mining and modeling prediction to explore the key factors affecting the level of enterprise technology innovation. It provides a scientific and quantitative method for evaluating the innovation capability of enterprises. Based on the enterprise's basic information and intellectual property-related data, comprehensive data preprocessing and feature engineering were conducted to screen out feature variables closely related to enterprise ratings. The Gradient Boosting Decision Tree (GBDT) was employed as the core model for regression prediction of enterprise ratings, demonstrating excellent performance in various error metrics (MSE, RMSE, MAE) and coefficient of determination (R^2), with small prediction errors and strong robustness, indicating the model's high adaptability in handling SME innovation level rating tasks.

In terms of data visualization, Python libraries (such as Plotly and Seaborn) and FineBI tools were used for multi-dimensional statistical analysis and intuitive display of enterprise data. The study specifically analyzed the association between basic information, intellectual property indicators, and enterprise innovation ratings, revealing the significant impact of features such as the number of effective invention patents and software copyrights on ratings. By analyzing regional distribution, feature importance, and the distribution of relative error and residuals in rating predictions, the study further validated the scientificity and explanatory power of the data and model, providing important decision-making references for policymakers, business managers, and industry researchers.

To enhance the practical application value of the research findings, an interactive Web interface was designed based on the Streamlit framework, enabling convenient prediction of enterprise ratings. Users only need to input relevant information such as the number of intellectual properties and enterprise type to obtain the model-predicted rating results. This interface not only provides a convenient prediction service but also visually presents the feature importance analysis for enterprise ratings, offering practical references for enterprises to understand their strengths and areas for improvement.

The innovation of this study lies in the combination of basic information and intellectual property indicators to construct a rating prediction model suitable for "specialized, refined, unique, and innovative" SMEs, and the practical transformation of model outcomes through a Web application. The research results not only provide technical support for the precise evaluation of enterprise innovation capabilities but also offer new insights for future industry trend analysis, policy optimization, and enterprise innovation path planning.

Keywords: Specialized, Refined, Unique, and Innovative SMEs, Data Analysis, Intellectual Property, Gradient Boosting Decision Tree

目录

1 绪论

1.1 研究背景和意义

1.1.1 现状与发展趋势

在全球经济进入新一轮科技革命和产业变革的关键时期，科技创新已成为提升国家竞争力和实现经济社会可持续发展的核心驱动力。中国政府高度重视中小企业的创新发展，早在2011年，工业和信息化部首次提出了培育“专精特新”中小企业的战略，旨在支持一批专业化、精细化、特色化、新颖化的中小企业成长壮大。

进入“十四五”时期，国家进一步加大了对“专精特新”中小企业的扶持力度。2021年7月30日，中共中央政治局召开会议，强调要发展“专精特新”中小企业。同年，工业和信息化部发布了《“十四五”中小企业发展规划》，提出到2025年，培育100万家创新型中小企业、10万家“专精特新”中小企业和1万家“专精特新”小巨人企业的目标。

广东省作为中国经济最具活力的地区之一，高度重视“专精特新”中小企业的培育工作。根据广东省工业和信息化厅的数据，截至2023年，全省已累计培育“专精特新”中小企业超过2.7万家，创新型中小企业超4.2万家，其中有近1528家企业被评为国家级“专精特新”小巨人企业。这些企业在推动区域经济发展、科技创新和产业升级中发挥着重要作用。

但是，许多中小企业在申请“专精特新”认定过程中仍面临诸多挑战，如：

1. 政策理解不足：企业对“专精特新”认定政策、评审标准和流程缺乏深入理解，导致申报材料准备不充分，影响认定结果。

2. 信息获取困难：缺乏便捷的渠道获取最新的政策动态、行业资讯和同行业标杆企业的信息，无法及时调整发展策略。

3. 专业指导缺乏：中小企业在科技创新和管理能力方面存在短板，缺少专业的咨询和指导，影响企业长远发展。

同时，政府部门在政策宣传和企业服务方面也存在信息传递效率低、服务精准度不足等问题，制约了政策红利的充分释放。

鉴于上述背景，亟需一个集政策咨询、企业自我评估和行业数据展示于一体的综合性服务平台，利用人工智能和大数据技术，提升企业对“专精特新”认定的理解和准备水平，助力中小企业的创新发展。

1.1.2 研究目的与意义

随着科技创新成为推动经济发展和增强国家竞争力的重要引擎，如何科学、全面地评价企业的科技创新能力，已成为社会各界关注的焦点。企业科技创新能力的定量化评估不仅能够为企业自我诊断提供参考，还能为政策制定者在资源分配、产业规划及创新支持等方面提供重要依据。但是传统的评分方法往往依赖专家打分或单一指标分析，存在主观性强、覆盖面不足、难以量化综合效应等局限性。因此，需要一种基于数据驱动的科学方法来改进企业科技创新评分体系的准确性和客观性。

本文的研究目的是基于机器学习技术，结合企业的多维度特征数据（如专利数量、研发投入、行业特性等），构建一个能够对企业科技创新评分进行高效预测的模型。通过引入梯度提升树算法，本文不仅力求提高评分模型的预测精度，还通过特征重要性分析揭示不同指标对企业科技创新能力的影响程度，提供更为透明的决策依据。

本研究的意义体现在以下几个方面：

1. 丰富了企业科技创新能力评价的理论体系，通过数据驱动的方法探索了关键指标的作用机理，为相关领域的学术研究提供了新视角和新方法。

2. 研究成果为政策制定者优化资源配置、制定扶持政策提供了科学依据，有助于推动企业科技创新能力的提升。同时，企业可以基于模型提供的评分结果和特征分析，有针对性地改善不足，提高市场竞争力。

3. 通过机器学习算法的应用，本研究展示了数据挖掘技术在企业评价领域的实际价值，为类似问题的解决提供了方法论参考。

本文以提升企业科技创新评分的科学性和应用价值为目标，致力于为促进创新驱动发展提供理论支持和技术工具。这一研究既服务于国家创新战略，又

有助于企业在全球竞争格局中占据更有利的位置，具有重要的理论和实际意义。

1.2 国内外研究现状

1.2.1 国内外研究成果与趋势

近年来，“专精特新”中小企业作为推动经济高质量发展的重要力量，得到了国内外广泛关注和研究。 国外对中小企业创新能力的研究主要集中在企业专业化、精细化、特性化和创新能力提升四个方面：

1. 专业化与创新路径研究：Cui 等（2017）提出，中小企业创新能力提升与社会企业创新扩散有着密切联系，强调技术创新在企业成长中的重要作用[12]。

2. 地理位置与政策环境的影响：Yi 等（2016）研究了地理位置和环境法规对企业创新转型的影响，指出区域环境和政策导向对企业创新能力提升至关重要[13]。

3. 绿色技术与数字化转型：刘新荷（2024）指出，在供应链金融的支持下，数字化技术对企业绿色技术创新具有显著促进作用，特别是对中小企业[10]。

国内“专精特新”企业研究主要围绕发展路径、政策支持与高质量发展展开：

1. 发展路径研究：刘昌年和梅强（2015）提出“专精特新”企业成长的多样化路径，强调资源整合和行业细分的重要性[3]。李培恩（2011）指出，中小企业通过专注于专业领域和技术突破，能够实现竞争力的提升[2]。

2. 政策支持的作用：财政部等部门发布政策支持“专精特新”中小企业高质量发展（佚名，2021），政策扶持成为推动企业创新发展的重要动力[1]。

3. 典型案例分析：朱宏任（2024）研究了中国企业从“专精特新”向世界一流企业发展的过程，探讨了企业家精神的关键作用[6]。

整体上，国内外研究在关注“专精特新”中小企业创新能力提升的基础上，逐步向以下趋势发展：评价体系更加全面、分析方法数据化和智能化、研究视角由单一维度向综合系统化转变。

1.2.2 现有研究的不足与本文研究重点

尽管国内外在“专精特新”中小企业创新能力研究中取得了丰硕成果，但仍存在以下不足之处：

1. 现有评价体系多依赖于少数核心指标（如研发投入、专利数量），未能充分考虑多维度特征，如企业的行业特性、区域影响和政策支持等，这可能导致评价结果存在局限性。

2. 大量研究采用机器学习等数据驱动方法提高预测精度，但部分模型缺乏可解释性，难以为政策制定和企业管理提供明确指导。

3. 现有研究多从整体角度探讨“专精特新”企业的创新能力，忽视了区域差异（如经济发达地区与欠发达地区）及行业特点（如制造业与服务业）的影响。

基于以上不足，本文的研究重点包括：

1. 构建多维创新能力评估体系：综合研发投入、专利成果、行业特性和政策支持等多维度数据，完善“专精特新”企业的创新能力评价模型。

2. 提升模型的解释性与实用性：采用机器学习中集成学习技术构建创新能力预测模型，并通过特征重要性分析揭示核心指标对创新能力的具体影响。

3. 探讨区域与行业差异：研究不同区域和行业背景对“专精特新”企业创新能力的影响，为区域政策制定提供数据支持。

本文力图弥补现有研究不足，为“专精特新”中小企业的高质量发展提供科学依据和实践指导。

1.3 本文研究内容与方法

1.3.1 研究内容概述

本文围绕“专精特新”中小企业的创新能力评价与预测展开研究，旨在构建科学、全面的评价体系，并开发一个可供企业及管理者使用的创新能力预测与分析平台。具体研究内容包括：

1. 多维度创新能力评价体系的构建结合中小企业的多维特征数据（如资本规模、专利成果、研发投入、行业类别等），构建全面的创新能力评价体系，为企业创新能力的量化研究奠定基础。

2. 创新能力预测模型的设计基于机器学习技术，采用梯度提升树算法，开发高效的创新能力预测模型。模型以历史数据为训练基础，能够预测企业未来的创新能力表现，并揭示核心影响因素的作用。

3. 行业数据的可视化分析运用多种可视化工具对行业和区域数据进行分析与展示，包括专利分布、研发投入趋势、区域行业分布等，帮助挖掘数据背后的深层次规律，为政策制定和企业管理提供参考。

4. 创新能力分析与预测平台的开发基于 Python 和 Streamlit 框架，开发一个便捷的 Web 应用。用户可通过输入企业相关数据，获得企业创新能力的预测结果和关键指标的分析。该平台支持快速交互，满足企业和政策研究人员的实际需求。

1.3.2 技术与方法介绍

为实现本文研究目标，综合采用了以下关键技术与方法：

1. 机器学习技术本文的预测模块基于梯度提升树（Gradient Boosting Decision Tree, GBDT）算法构建。GBDT 是一种高效的集成学习方法，能够通过逐步优化残差的方式构建强预测模型。GBDT 具有高预测精度和对特征重要性进行解释的能力，适合多维度特征的企业创新能力预测场景。

2. 数据可视化技术本文在行业数据分析与展示中，结合 Python 的可视化库（如 Matplotlib、Seaborn）与 FineBI 工具。Python 库用于生成高度定制化的图表，如折线图、散点图等；FineBI 则支持快速构建业务仪表板，直观展示行业和区域的综合数据指标。

3. Web 前端开发技术采用 Streamlit 框架开发预测平台。Streamlit 是一种轻量化的 Python Web 开发工具，支持快速构建交互式数据科学应用。结合其实时数据更新能力，用户能够方便地使用预测功能并浏览动态分析结果。

4. 开发工具与环境

选用 Python 3 作为主要编程语言，其丰富的生态系统能够满足数据分析、机器学习和 Web 开发等多种需求。以及 PyCharm 作为集成开发环境（IDE），其高效的调试和插件支持功能，为快速开发与测试提供了便捷条件。利用 FineBI 作为行业数据分析和展示的主要平台，通过其丰富的数据处理功能和交互式仪表板，进一步提升了行业数据的可视化效果。

通过上述技术的整合，本文在理论模型与实践应用之间搭建了桥梁，为“专精特新”中小企业的创新能力分析提供了科学支持和技术保障。

1.4 论文基本结构

本文围绕“专精特新”中小企业的评分与分析展开研究，整体内容包括数据处理、探索性分析、模型构建及预测，并开发交互式 Web 界面以展示研究成果和提供预测服务。各章节具体内容如下：

第一章：绪论阐述研究背景及其在国家政策支持和中小企业高质量发展中的重要意义。结合国内外研究现状，梳理了“专精特新”企业评价领域的主要成果与不足，明确了本研究的方向和价值，最后概括性描述论文结构，为全文奠定理论与实践基础。

第二章：相关技术介绍本章介绍了数据分析与建模所需的相关技术与方法。包括：

- 数据可视化技术与工具的基本概念，重点介绍 Matplotlib、Seaborn 和 FineBI 的应用；

- 机器学习理论及算法原理，涵盖梯度提升树模型及其在回归问题中的优势；
- 数据预处理和特征工程技术，为模型的准确性和稳定性提供支持。

第三章：数据介绍与处理本章详细说明了研究数据的来源与特征，结合任务需求对数据进行了系统化的处理，包括缺失值填补、异常值剔除、数据标准化和特征编码等操作，确保数据的质量与一致性，为后续模型构建奠定基础。

第四章：数据探索与分析本章以探索性数据分析（EDA）为核心，利用 Python 可视化库（如 Matplotlib 和 Seaborn）及 FineBI 对数据进行多维度解析，揭示评分与基础信息、知识产权等特征的潜在关系。同时，利用统计分析方法深入挖掘了影响企业评分的关键因素，为模型训练提供了依据。

第五章：模型构建与预测本章分为四部分：

1. 特征工程：基于数据特征分析结果，构造并筛选影响评分的核心变量；
2. 模型训练与评估：选用梯度提升树算法训练评分预测模型，并通过 MSE、R^2 等指标评估模型性能；
3. 模型选择与预测：对比多种模型的优劣，最终选择最优模型进行企业评分预测；
4. Web 应用开发：基于 Streamlit 框架，设计交互式 Web 界面，展示预测结果并支持用户输入企业信息进行评分预测。

第六章：总结与展望本章总结研究成果与创新点，指出评分预测模型和 Web 应用在实践中的价值。同时，分析了研究的不足之处，并提出未来研究方向，包括扩展数据维度、优化模型性能、开发更多功能模块等。展望中进一步强调本研究的潜在应用价值，如政策支持、投资分析及企业管理等领域的智能化决策支持。

第七章：参考文献汇总研究中参考的学术文献和技术资料，涵盖数据分析、机器学习、可视化技术和企业发展相关研究，为本文内容提供权威依据。

第八章：致谢本章对研究过程中提供指导与支持的导师、团队以及所有帮助完成本研究的人员表示感谢。

2 相关技术介绍

2.1 数据可视化

数据可视化是通过图表、图形等方式对数据进行直观展示的技术方法。其目标是简化数据理解、揭示隐藏规律并辅助决策。常见的方法包括静态图表（如折线图、柱状图）和动态仪表板（如数据交互平台）。主要工具包括Python数据可视化库（如Matplotlib、Seaborn）和商业可视化平台（如FineBI、Tableau）。

部分使用到的可视化库：

1. Matplotlib：功能强大，支持多种图表类型，如折线图、散点图、直方图等，适合高度定制化需求，常用于生成学术级和出版级的静态图表，但美观度一般。

2. Seaborn：基于Matplotlib，注重统计数据可视化，支持更高级的图表类型，如热力图、箱线图和分布图，同时优化了图表的美观性和易用性。

3. Plotly：交互式可视化库，支持生成动态、可交互的图表，如3D图表、地图和嵌套图。它适合在数据分析报告和Web应用中展示复杂的数据关系。

可视化库的综合应用，为数据分析、模型结果展示以及Web界面开发提供了全面支持。

2.2 机器学习

机器学习是一种让计算机从数据中学习规律并预测结果的技术，广泛应用于分类、回归和聚类等任务。机器学习的原理主要分为监督学习、非监督学习和强化学习三大类。其中，监督学习以带标签的数据为基础，通过训练模型实现预测目标；非监督学习用于发现数据内部的潜在结构；强化学习则模拟智能体在特定环境中优化决策过程。

部分模型介绍：

1. 梯度提升树（GBDT）：一种集成学习方法，通过逐步优化弱学习器的预测误差提高模型性能。GBDT 在特征选择和预测精度方面表现优秀，适合多维数据的分析与预测。

2. 逻辑回归：广泛用于分类问题的线性模型，适合快速训练和解释数据。

3. 随机森林：基于多棵决策树的集成方法，能够有效处理非线性关系并避免过拟合。

4. 支持向量机（SVM）：一种基于边界最大化的分类模型，适合高维数据分析。

2.3 其他相关技术

数据预处理是机器学习的重要步骤，确保数据质量和模型效果。主要工作包括数据清洗、缺失值处理、异常值检测、数据标准化和归一化等。 其中数据清洗主要为修正错误数据，填补缺失值（如均值填充、插值等）。标准化和归一化即为将数据转换为统一的尺度，避免特征间尺度差异影响模型性能。

特征工程是通过设计和优化特征提高模型性能的过程。主要方法包括：

1. 特征选择：剔除冗余特征，减少数据维度。常用方法包括基于统计量的筛选和基于模型的特征重要性评估。

2. 特征转换：通过对原始数据进行变换（如对数变换、分箱处理）改善特征分布。

3. 特征编码：将类别型变量转化为数值型表示，常见方法有独热编码（One-Hot Encoding）和标签编码（Label Encoding）。

通过结合以上技术与方法，本文为“专精特新”中小企业创新能力的科学评价和高效预测打下基础。

3 数据介绍与处理

3.1 数据来源与说明

本系统的数据来源主要包括三个渠道：一是通过广东省工业和信息化厅获取了 2023 年广东省专精特新中小企业的名录；二是利用天眼查平台，搜集企业的基本信息、知识产权和科技创新相关指标数据；三是百度开放平台 api 获取企业地理位置经纬度信息。

在 2023 年广东省专精特新中小企业的名录中，总共包含了 5970 个企业。根据此名单成功在天眼查导出所有数据，总计 17 个变量，包含公司名称、登记状态、注册地址、注册资本等信息。

通过基础信息中“注册地址”数据，利用百度地图 api 获取地址对应的经纬度信息。

因企业知识产权与科创指标信息无法直接导出，采用自动化软件采集的方式进行获取，成功获取到 5956 条数据。

3.1.1 数据集概述

由于原始表字段命名不统一，且部分字段数据对该研究无意义需要删除、部分字段需要进行再次处理，故展示整理后的字段、含义及说明，如表 3.1：

表 3.1 字段及含义

字段名	中文字段名	含义
id	编号	企业的唯一标识符
company_name	企业名称	企业的正式名称
address	注册地址	企业注册时的法定地址
longitude	经度	地址对应的地理坐标中的经度值
latitude	纬度	地址对应的地理坐标中的纬度值
capital	注册资本	企业在注册时申报的资金总额
capital_cny_w	注册资本（单位：万元）	注册资本的金额以万元为单位表示
founding_date	成立日期	企业成立的日期

字段名	中文字段名	含义
city_district	所在城市和区	企业所在的行政区域
type	企业类型	根据法律或行业标准划分的企业类别
scale	企业规模	按照员工数量、资产规模等标准评估的企业大小
insured_number	参保人数	为企业提供社会保险的人数
grade	科创水平级别	对企业的信用等级或其他评价标准的评定结果
score	科创水平分数	量化分析企业科创水平高低的指标；取值 0-100
industry_lv2	二级行业分类	行业分类体系中的二级分类
exceed_percent	超越同行百分比	表示企业超过同行的比例
invention_patent_applications	发明专利申请数	企业提交的发明专利申请的数量
valid_invention_patents	有效发明专利数	当前有效的发明专利数量
valid_utility_models	有效实用新型专利数	当前有效的实用新型专利数量
valid_design_patents	有效外观设计专利数	当前有效的工业品外观设计专利数量
software_copyrights	软件著作权数	企业拥有的软件著作权数量
innovation_capability	创新能力	衡量企业创新能力的综合指标；取值 0-20
r_d_strength	研发实力	衡量企业研发能力的综合指标；取值 0-20
industry_potential	行业潜力	衡量企业行业潜力的综合指标；取值 0-20
growth	成长性	衡量企业成长性的综合指标；取值 0-20

字段名	中文字段名	含义
tech_innovation_qualification	科创资质	衡量企业科创能力的综合指标；取值 0-20
matched_certification_list	匹配的认证列表	企业获得的所有认定资质
business_scope	经营范围	企业被允许从事的业务活动领域

表 basic.xlsx 数据概览，在下文简称表一，如图 3.1、3.2：

id	公司名称	登记状态	注册地址	注册资本	成立日期	所属省份	所属城市	所属区县	公司类型
1	广州第七轴智能设备有限公司	开业	广州市南沙区大涌路62号	500万人民币	2020-08-17	广东省	广州市	南沙区	有限责任公司(自然人投资或控股)
2	广州电玩时代科技有限公司	开业	广州市天河区莲溪路4号	100万人民币	2017-12-29	广东省	广州市	天河区	有限责任公司(自然人投资或控股)
3	广州顶盛益电子科技有限公司	开业	广州市增城区新塘镇新华	20万人民币	2017-03-03	广东省	广州市	增城区	有限责任公司(自然人投资或控股)
4	广州鼎展汽车零部件有限公司	开业	广州市花都区新华街马溪	1000万人民币	2013-09-27	广东省	广州市	花都区	有限责任公司(自然人投资或控股)
5	广州鼎得康高分子材料有限公司	开业	广州市从化区太平镇沛益	100万人民币	2015-12-07	广东省	广州市	从化区	其他有限责任公司

图 3.1 表 basic.xlsx 数据概览 1

国标行业门类	国标行业大类	国标行业中类	企业规模	实缴资本	经营范围
制造业	计算机、通信和其他电子设备制造业	电子器件制造	中高型	-	专用设备制造（不含许可类专业设备制造
科学研究和技术服务业	科技推广和应用服务业	科技中介服务	中型	-	科技中介服务;运行效能评估服务;机械设
科学研究和技术服务业	研究和试验发展	工程和技术研究和试验发展	微型	-	电子、通信与自动控制技术研究、开发;电
制造业	橡胶和塑料制品业	塑料制品业	大型	100万人民币	塑料零件制造;塑料保护膜制造;塑料包装
制造业	化学原料和化学制品制造业	涂料、油墨、颜料及类似产品制造	中型	100万人民币	颜料制造;粘合剂制造（监控化学品、危险

图 3.2 表 basic.xlsx 数据概览 2

该数据集包含了关于不同专精特新中小企业的详细信息。包括了公司名称、注册地址、成立日期等基本信息；同时包括行业分类、企业规模以及经营范围等更深入的信息。有助于了解每家公司的成立时间、规模、行业与业务范围。

表 kechuang.xlsx 数据概览，在下文简称表二，如图 3.3：

company_name	grade	score	industry_lv2	ceed_perce	patent_ap	nvention_pi	utility_mo	design_pa	vare_copyri	five_score	matched_certification_list
佛山峰合精密喷射成形科技有限公司	优秀	81	研究和试验发展	90%	16	10	4	0	2	13,12,13,10,13	高新技术企业;专精特新企业;科技型
广东省建科建筑设计院有限公司	优秀	85	专业技术服务业	93%	64	19	54	0	8	14,12,14,10,13	高新技术企业;专精特新企业;
侨益物流股份有限公司	卓越	95	多式联运和运输代理业	99%	10	8	7	0	23	16,14,10,18,9	专精特新企业;瞪羚企业;小微企业;创
广东建筑艺术设计院有限公司	良好	74	专业技术服务业	84%	0	0	8	0	40	11,12,14,10,13	高新技术企业;专精特新企业;科技型
新安天玉有机硅有限公司	优秀	81	化学原料和化学制品制造业	85%	37	22	33	0	0	12,12,11,10,16	高新技术企业;专精特新企业;企业技

图 3.3 表 kechuang.xlsx 数据概览

该数据集展示了不同专精特新中小企业在创新能力和研发实力方面的数据。其中包括公司的评分、所属行业、超过百分比、专利申请情况（包括发明、实用新型和外观设计）、软件著作权数量以及创新能力等指标。有助于了解各个公司在技术创新和发展潜力方面的表现，为该研究主要的数据内容。

表 Latitude_longitude.xlsx 数据概览，在下文简称表三，如图 3.4：

公司名称	登记状态	注册地址	经度	纬度
广州第七轴智能设备有限公司	开业	广州市南沙区大涌路62号	113.5517	22.76453
广州电玩时代科技有限公司	开业	广州市天河区莲溪路4号4	113.4191	23.10601
广州顶盛益电子科技有限公司	开业	广州市增城区新塘镇新何	113.6156	23.12051
广州鼎晟汽车零部件有限公司	开业	广州市花都区新华街马溪	113.1502	23.3548
广州鼎得康高分子材料有限公司	开业	广州市从化区太平镇沛益	flase	flase

图 3.4 表 Latitude_longitude.xlsx 数据概览

该数据集主要为企业的地理位置信息，具体包含了各专精特新中小企业的经纬度数据。为后续的地理空间分析提供了基础，探究不同企业在区域分布上的选择偏好及其异同。

3.1.2 数据特征分析

数据集中部分字段的取值数据，如表 3.2：

表 3.1 部分字段描述性统计表

统计量	注册资本（万元）	参保人数	科创水平分数	超越同行百分比	发明专利申请数	有效发明专利数	有效实用新型专利数	有效外观设计专利数	软件著作权数
count	5956	5956	5956	5956	5956	5956	5956	5956	5956
mean	3565.846	89.227	78.805	83.865	9.062	3.523	16.662	3.927	8.573
std	1846.955	188.518	10.856	13.307	34.595	18.612	21.129	14.138	21.208
min	3	0	46	1	0	0	0	0	0
25%	500	19	72	77	1	1	6	0	0
50%	1000	41	78	87	4	1	13	0	0
75%	3000	91	86	94	8	3	22	2	8
max	1200000	8207	99	99	2022	1229	471	294	395

数据集包涵了 5970 家企业的参保人数、科创水平分数、超越同行百分比、发明专利申请、有效发明专利、有效实用新型专利、有效外观设计专利以及软件著作权等关键指标。

结果显示，注册资本均值为三千五百万元左右，体现获得该认定的企业有一定规模；平均参保人数为89.19人，但标准差高达188.40，表明企业规模各有差异；科创水平分数平均值为78.80，标准差为10.86，显示出科创水平大多在70-90区间。超越同行百分比的平均值为83.87%，反映出大多数企业在行业中处于较高水平。在专利申请和持有方面，各单位展现出显著的技术创新能力差异，发明专利申请的平均数为9.06项，而有效发明专利的平均数为3.52项，实用新型和外观设计专利的持有量也呈现出较大的离散性。此外，软件著作权的平均数为8.57项，揭示了企业在软件开发和知识产权保护方面的投入差异。该数据揭示了企业在规模、绩效、技术创新和知识产权保护等方面的差异，为深入探究企业发展的驱动因素提供了数据支持。

3.2 数据预处理

在数据采集与记录的初步阶段，识别出数据集中存在若干质量问题，包括缺失值、异常值等，这些问题对于数据分析的严谨性至关重要。如忽视这些问题的处理，分析结果会产生一定程度误差，进而影响结论的准确性和可信度。因此在进行数据分析之前，须对数据进行彻底的清洗和预处理，以确保数据集的质量和分析结果的可靠性。

具体到各个数据表，问题如下：

表一：该表面临的主要问题包括：部分字段存在高达20%的缺失值，这一比例的缺失数据将对后续分析产生影响；字段命名缺乏规范性，这可能导致在数据处理和分析过程中产生混淆；最后，一些关键字段的数据存在异常格式，这会对数据的整合和分析造成障碍。

表二：在此表中，"超越同行百分比"字段目前为字符串类型，包含百分号，需要将其转换为数值类型，以便进行数学运算。同时，科创指标数据需要进行拆分，以便更细致地分析各个科创维度，为后续的数据分析工作打下坚础。

表三：在地理信息可视化过程中，发现少数经纬度数据不在广东省范围内，将这部分归为异常数据。需要对这些异常坐标进行筛选并删除，以确保地理分析的相关性和准确性。

并且考虑到后续编码过程中的便捷性与易用性，将三个数据表根据"企业名称"这一共同字段进行合并，并将所有字段名称转换为英文字段名称，以防止在数据库建立与查询过程中出现字符编码错误。

3.2.1 数据清洗与处理

在数据清洗与处理阶段，采取了多项措施以确保数据的质量和分析的准确性。首先将数据整理为对应的类型格式，以及对"注册资本"进行了汇率转换，以 2023 年的汇率为准，将港元、美元、欧元、日元和加元对人民币的汇率进行转换，从而统一了数据的单位。并对地理坐标点进行了处理，定义了广东省的地理坐标范围，并检查每个点是否在该范围内。对于不在范围内的坐标点，将其对应的 ID 保存下来，并在后续步骤中将其坐标设置为 NaN，以避免在地理分析中出现错误。最后对该表的字段名进行重命名为英文字段名。

以下为表一较多缺失值的字段及缺失值数量统计，如表 3.3：

表 2.3 部分字段缺失值统计表

表 1 字段名	缺失值数量
所属区县	1420
国标行业门类	1113
国标行业大类	1113
国标行业中类	1113
实缴资本	1021

其中“所属区县”字段有较多缺失值的原因为东莞市与中山市未设立行政区，导致数据为空，因此合并“所属城市”与“所属区县”字段，如广州市天河区或东莞市，为后续数据分析排除干扰。

而其中关于行业类型的数据有较多缺失，可能是由于天眼查中该企业在申报时未填写该字段导致，考虑到表二已有“industry_lv2”二级行业分类信息，且无缺失值，因此对表 1 的行业类型数据进行删除，方便后续整合数据集。

针对表二，对"five_score"列进行了拆分，将其拆分为五个单独的列，分别代表五个不同的评分指标，以便于后续的分析和可视化。

通过这些数据清洗和处理步骤，有效地提高了数据的质量和可用，为后续的分析工作做准备。

3.2.2 预处理目的与效果

数据分析的预处理环节，涉及数据的清洗、集成、转换和规约等多个方面，其核心宗旨在于保障数据集的质量、一致性与可用性，进而为后续的分析工作奠定基础。以下是预处理环节的几个主要目标，旨在提升数据分析的整体效能：

1. 数据质量提升：预处理过程中，通过对数据进行彻底的清洗与整理，能够有效识别并修正数据集中的错误信息、填补空白值，减少数据的不一致性。

2. 数据可用性增强：通过执行数据类型转换和跨源数据融合操作，预处理能够将分散在不同来源、具有多样格式的数据统一整合，提升了数据集的适用性。

3. 数据处理效率优化：通过精简字段和拆分数据结构，预处理有助于降低数据集的体积与复杂度，从而显著提高数据处理的效率，并加速分析流程的推进。

主要预处理步骤：

1. 找出数据中的缺失值格式，如“-”、“false”，替换为标准的pandas 的 Nan 格式，统计数量；

2. 删除无数据分析意义的字段及其数据；

3. 以“企业名称”合并 3 个表，由于表 1 存在曾用名、地名的问题，对其进行了正则处理；

4. 对合并后的数据集进行字段名称规范，将其中的中文字段名改为其英文映射；

5. 对“注册资本”字段进行数据单位统一，以 2023 年汇率为基准，将港元、美元、欧元等单位统一转为人民币；

6. 处理不在广东省内的异常坐标点，先确定广东省边缘的多个坐标点，利用 shapely 库的 ploygon 方法将坐标点设置为多边形，后遍历数据判断是否位于多边形内部，否则标记为"false";

7. 拆分科创指标数据为创新实力、研发能力等字段；

8. 对字段顺序进行排列，提高数据集易读性。

预处理后字段缺失值统计如下，如表 3.4：

表 3.3 预处理后表 1 缺失值统计表

字段名	缺失值数量	字段名	缺失值数量
id	0	industry_lv2	0
company_name	0	exceed_percent	0
address	0	invention_patent_applications	0
longitude	21	valid_invention_patents	0
latitude	21	valid_utility_models	0
capital	0	valid_design_patents	0
capital_cny_w	0	software_copyrights	0
founding_date	0	innovation_capability	0
city_district	0	r_d_strength	0
type	0	industry_potential	0
scale	0	growth	0
insured_number	0	tech_innovation_qualification	0
grade	0	matched_certification_list	0
score	0	business_scope	0

在整理完成后的数据集中，字段格式得到统一，数据格式也进行较为完善的处理，绝大多数字段没有缺失值，只有经纬度信息存在极少量缺失值，在后续的地理位置分析过程中，影响程度较低，将缺失经纬度的数据不做展示。

4 数据探索与分析

在数据分析的可视化技术选择上，本研究主要采用 Python 进行基础可视化，了解数据的内在特征和规律。同时为了增强数据展示的交互性和可视化效果，进一步使用 FineBI 制作数据可视化大屏，供潜在的“专精特新中小企业”创业者查看。通过 FineBI，企业家能够更加直观地了解已获得此认定的企业特征及其分布情况。

在数据可视化结果展示过程中，本文将同时展示基于 Python 绘制的图表和使用 FineBI 制作的可视化图形。Python 可视化图表侧重于数据分析的结果和结论，重点揭示数据的模式和趋势；而 FineBI 图表则侧重于图形展示的效果和用户交互体验，以提高图表的可操作性和可视化的精细度，为用户提供更为便捷的在线查看和分析工具。

其中 python 利用可视化库进行，主要为 Matplotlib、Seaborn、Plotly 库。

4.1 数据探索性分析（EDA）

数据探索性分析（Exploratory Data Analysis，简称 EDA）是数据分析的重要步骤，主要目的是了解数据的基本情况，发现数据中的模式、关系和异常情况，为后续的模型建立提供依据。上文已完成了数据集的整理，缺失值统计与数据描述性统计分析，包括均值，中位数，标准差等。因此在下文将分为进行数据分布情况的基础可视化分析，与字段间相关性可视化分析，前者主要了解数据的频数特征，后者着重于字段之间的关系探索。

4.1.1 描述性统计与相关性分析

下文将展示企业规模、科创水平等级、所属行业、企业类型、所在区域的数据频数可视化并加以分析：

以下为企业规模方面的频数直方图，并与科创水平等级相比较，如图 4.1、4.2、4.3：

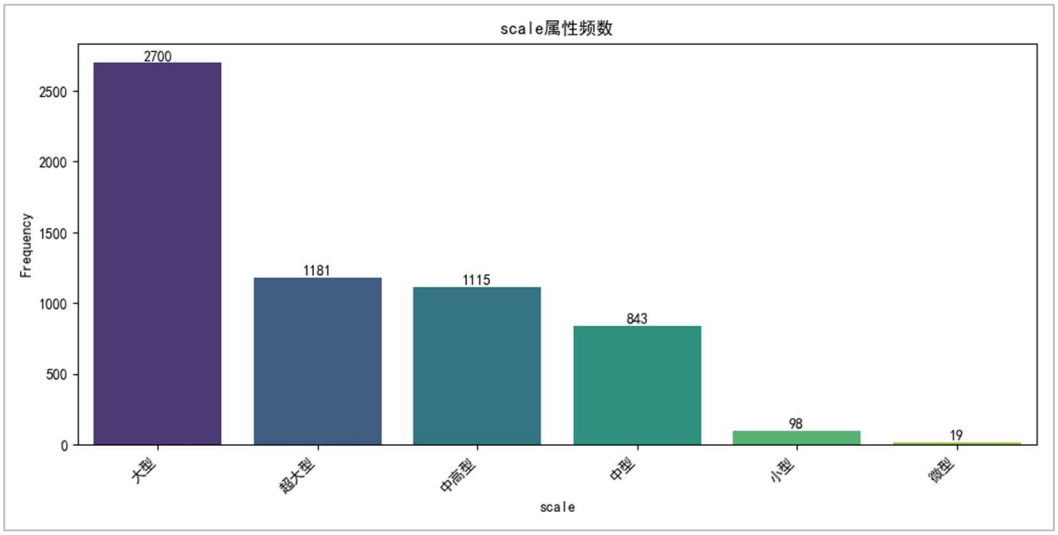

图 4.1 企业规模频数直方图

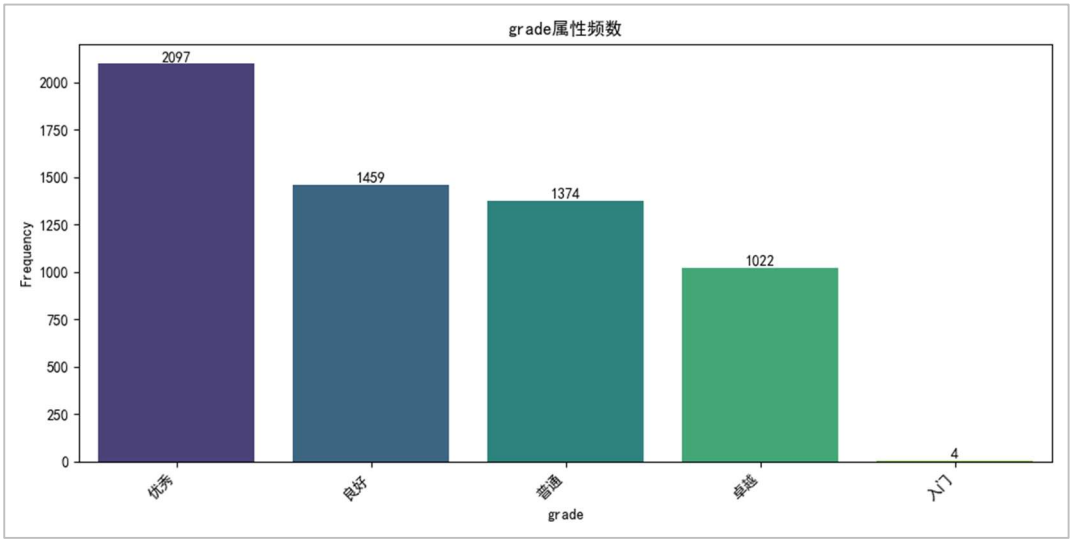

图 4.2 企业科创水平等级频数直方图

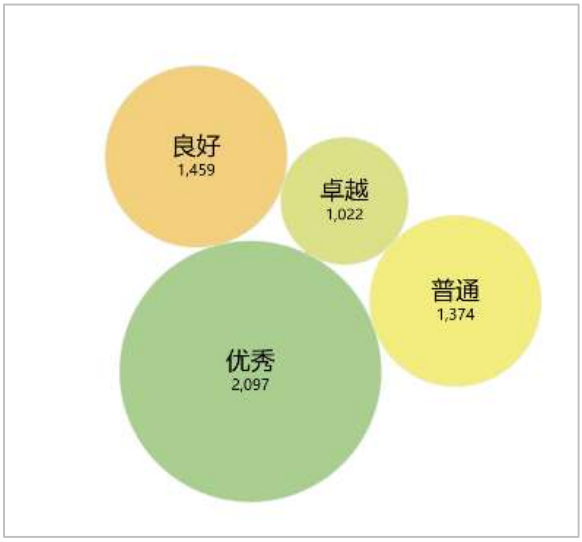

图 4.3 企业科创水平等级频数气泡图

根据上图，可等得出该数据集的企业在规模上多数为大型企业，占比近50%，并且科创水平处于”优秀“、”良好“的企业数量占比近60%，体现出获得”专精特新中小企业“认定的企业较为优秀。显示出该认定不仅在规模上有一定的要求（以大中型企业为主），而且在科技创新方面也有较高的标准和期望，意味着认定标准更倾向于具有一定规模和市场影响力的企业。但其中不乏少量中小微企业，体现了政策对于不同规模企业发展的全面考虑。

企业所属行业频数直方图与扇形图，如图4.4、4.5：

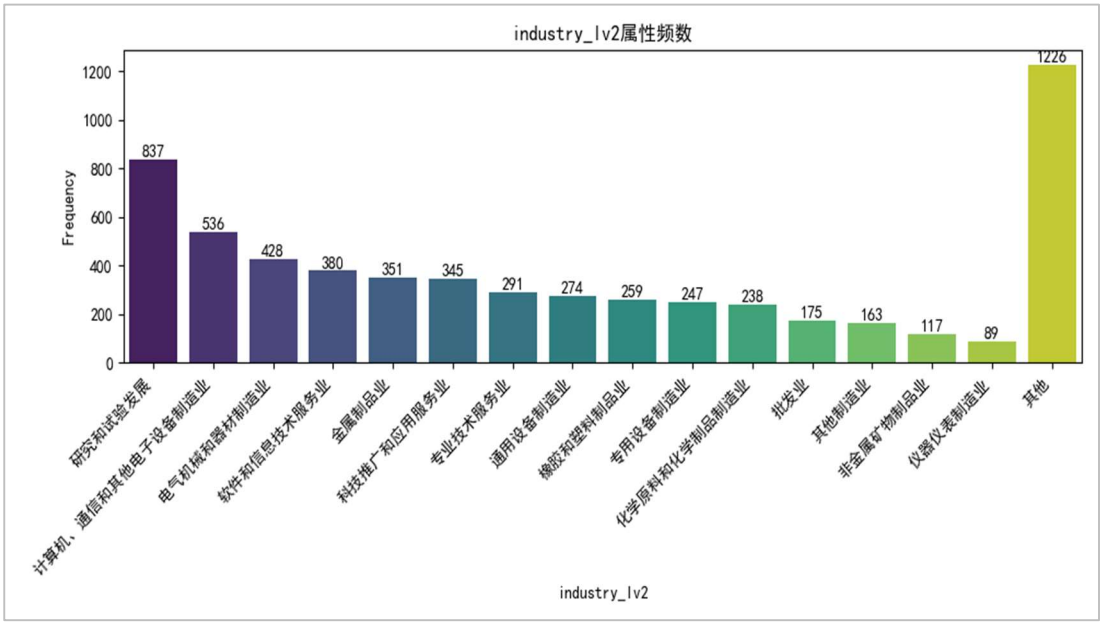

图 4.4 企业所属行业频数直方图

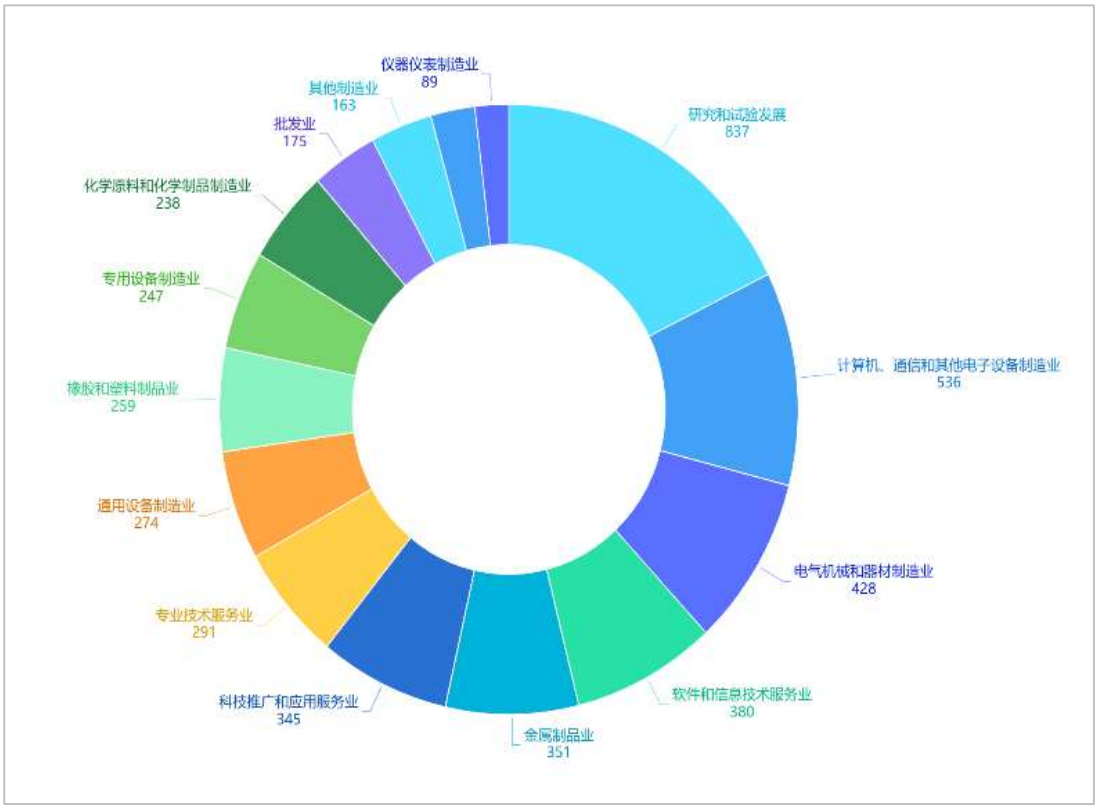

图 4.4 企业所属行业频数扇形图

可以看出企业所属行业分类比较广泛，其中”研究和试验发展“、”计算机、通信和其他电子设备制造业“、”电器机械和器材制造业“位居前三，一定程度上体现了科技创新和高新技术领域是比较受到重视的，也是推动地区经济发展和产业升级的关键，也反映出市场需求的增长和技术进步的趋势，特别是在数字化、智能化和绿色发展等新兴领域。政府的重视和支持，结合市场需求和技术发展，共同推动了这些行业的快速发展和企业数量的增加。

企业类型（法律责任结构）频数直方图，如图 4.6：

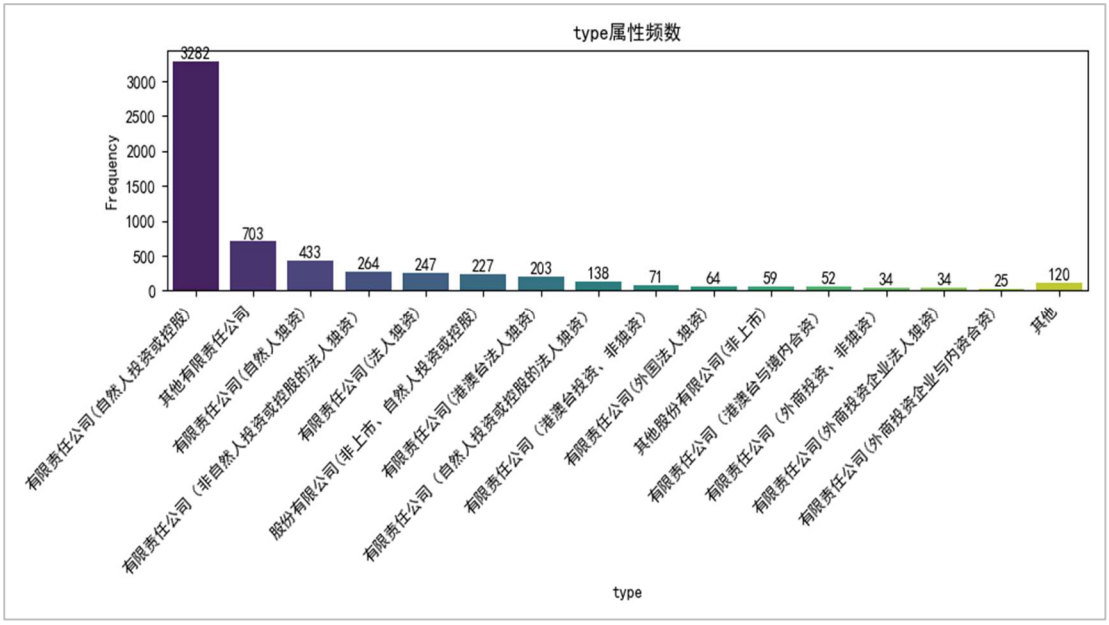

图 4.6 企业类型（法律责任结构）频数直方图

可以看出有限责任公司（自然人投资或控股）以 55.10%的比例占据主导地位，反映其在灵活性、责任限制、控制权、融资便利和税收优势等方面的综合优势。这些优势使得这种企业形式成为个人投资者的首选，尤其是在创新和专业化领域，个人投资者的直接参与可以更好地推动企业的发展和创新。而其广泛存在也与政府的政策导向有关，政府可能通过提供税收优惠、简化注册流程等措施来鼓励个人创业和投资。这种政策支持有助于激发市场活力，促进经济的多元化发展。

在专精特新中小企业中，自然人投资或控股的有限责任公司的主导地位，有助于推动技术创新和产业升级，也为个人投资者提供了实现价值和财富增长的平台。这种企业形式的普及，对于构建一个充满活力、创新驱动的经济体系具有重要意义。

以下为企业所在区域的频数直方图，如图 4.7：

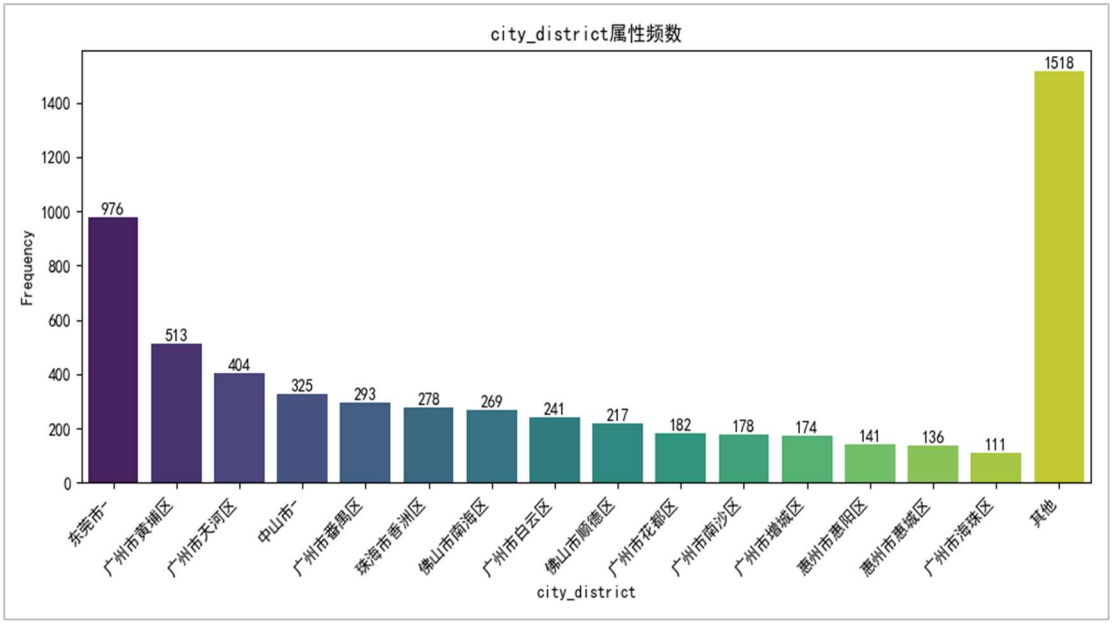

图 4.7 企业所在区域的频数直方图

该图表展示了“专精特新中小企业”在广东省的区域分布情况。由于广东省约有 110 个行政区，为了更清晰地观察重点分布区域，本图表筛选了企业数量排名前 15 的区域，其余区域归为“其他”类别。

从图中可以看出：

1. 东莞市位居第一，拥有 976 家企业，其没有具体划分为行政区，作为一个整体统计，因此整体企业数量相对集中。东莞市是广东省重要的制造业基地，以中小企业为主导，其成熟的产业链、完善的供应链配套以及政府对创新型企业的大力扶持，为“专精特新中小企业”发展提供了优越条件。

2. 广州市黄埔区和广州市天河区位居二、三名，分别为 513 家和 404 家，其为广州市创新经济和高科技产业的核心地带。其中黄埔区拥有广州开发区等高新技术园区，政府政策支持力度大，产业创新能力突出，吸引了大量企业。天河区则是广州市的商业中心，同时在信息技术和现代服务业领域具有优势，为科技型中小企业的发展提供了肥沃的土壤。

3. 广东省“专精特新中小企业”呈现显著的区域集中性，主要分布在珠三角核心城市的重点区域，如东莞、广州、佛山、中山，体现了这些区域发达的经济水平、完善的产业链体系以及政府的政策支持密不可分。

以下为地理位置可视化展示，按科创等级划分，如图 4.8：

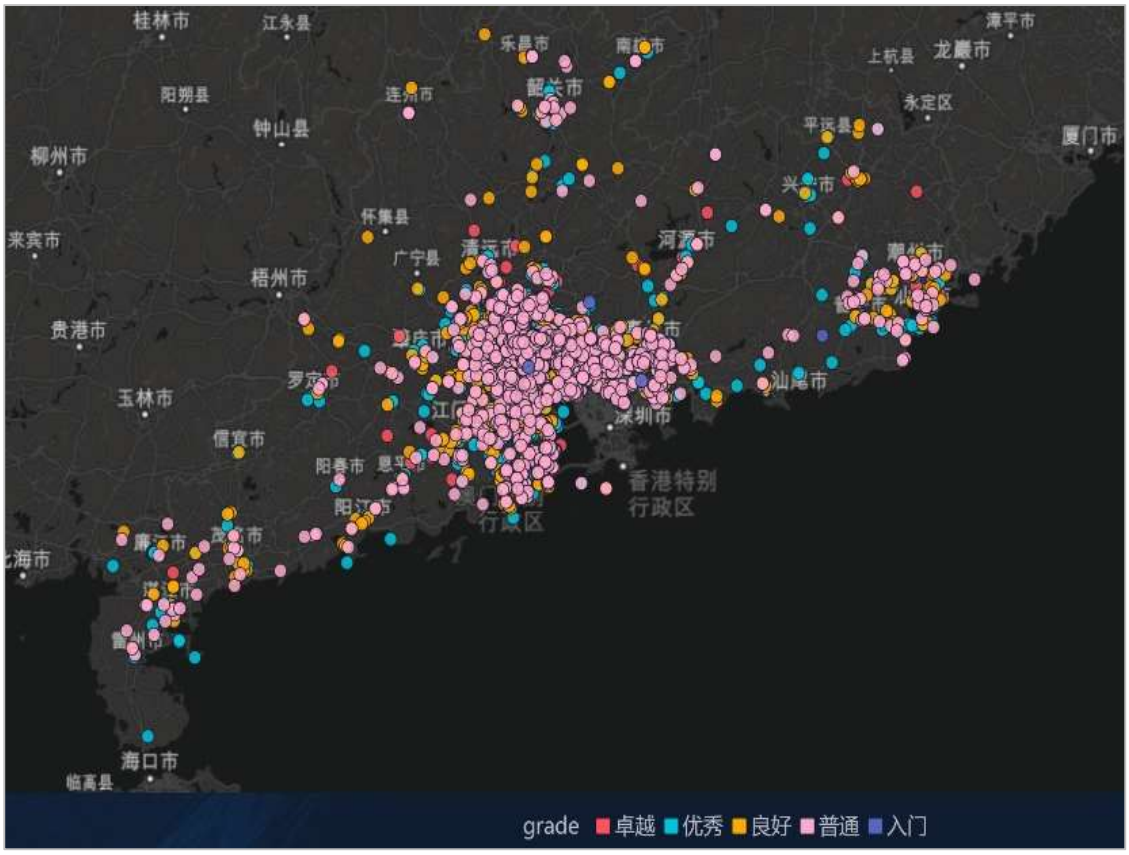

图 4.8 企业地理位置可视化图

图中可以看出：企业分布明显集中在珠三角区域，尤其是广州市、东莞市和佛山市等核心城市，这些区域的企业密度显著高于广东省其他地区。在珠三角区域中，不仅企业数量多，同时企业等级（如“卓越”“优秀”）的占比也较高，体现了这两地的产业集聚效应和高质量发展优势。究其原因，珠三角地区作为广东省的经济核心地带，具备完善的产业链、研发基础以及高效的政策支持，使得“专精特新中小企业”高度集中并表现出强劲的创新能力。相比之下，粤东和粤西地区则可能因基础设施、技术支持和政策扶持不足，导致企业分布较为分散，整体等级偏低。

以下为企业经营范围词频统计图与词频图，如图 4.9、4.10：

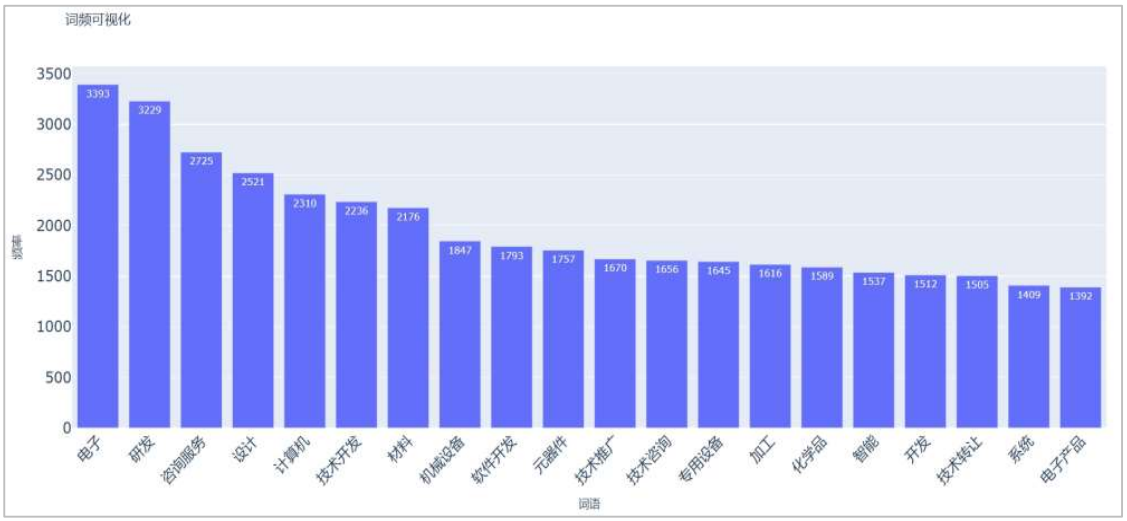

图 4.9 企业经营范围词频统计图

图 4.10 企业经营范围词频词云图

图中可以看出专精特新中小企业的经营重点集中于高科技和创新驱动领域，同时也包含传统制造业与专业服务的融合。从经营范围的高频关键词分析来看，“电子”、“研发”、“计算机”等体现了企业以科技创新为核心，推动数字化转型和技术开发的显著特征；而“材料”、“机械设备”等则表明制造业在区域经济中仍占重要地位，通过与科技融合实现转型升级。此外，“咨询服务”和“设计”等专业服务的高频出现，反映了企业在创新过程中对外部支持的依赖性，以及服务业在提升企业竞争力中的关键作用。体现出专精特新中小企业正通过科技创新与行业协同，实现多元化经营与高质量发展的平衡。

以下为企业成立时间（年份）频率折现图与各年增长率图，如图 4.11、4.12：

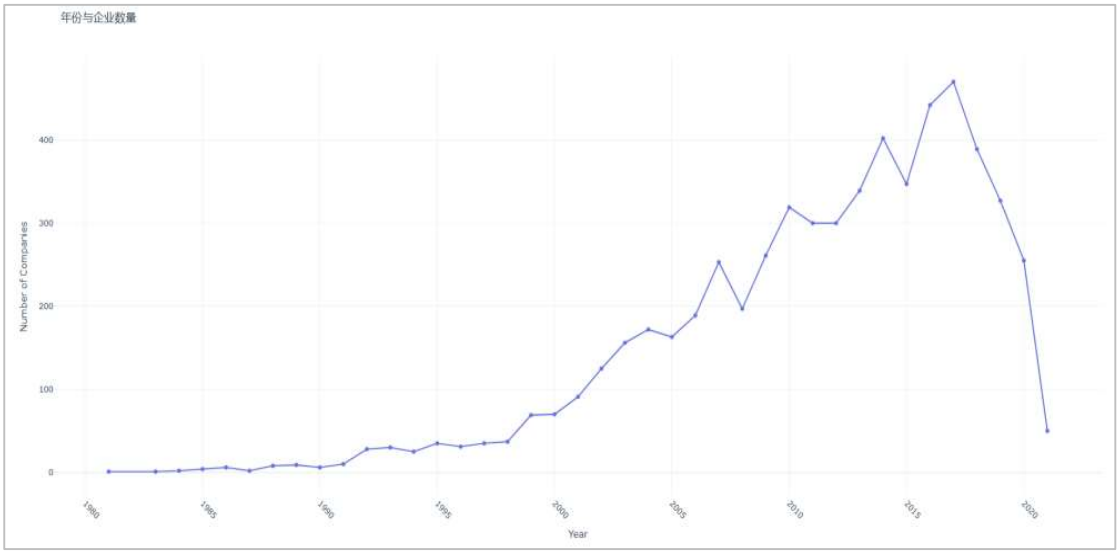

图 4.11 企业成立时间（年份）折线图

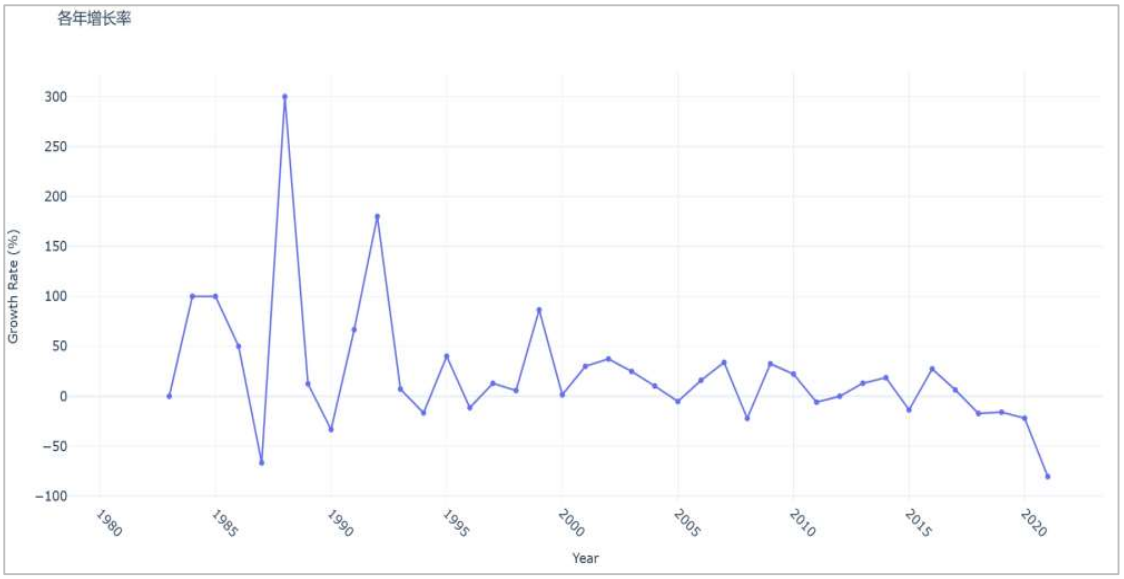

图 4.12 企业成立时间（年份）增长率折线图

专精特新中小企业的成立时间分布呈现出显著的阶段性特征。从 2000 年起，企业数量开始迅速增加，并在 2015 年前后达到顶峰。这一趋势与我国经济发展、产业政策以及时代技术的进步密切相关。2000 年到 2015 年是中国经济高速发展的重要阶段，我国加入 WTO 后，市场开放度和国际竞争力显著提高，为企业创新发展创造了良好的外部环境。同时，信息技术、互联网技术的普及

以及制造业升级需求，推动了技术型中小企业的快速崛起。尤其是 2015 年前后，国家大力推动“双创”（大众创业、万众创新）政策，并加大对科技型中小企业的资金和技术支持，使得这一时期成为专精特新中小企业成立的高峰阶段。

2020 年以后，企业成立数量出现下降趋势，可能与新冠疫情导致的经济环境不确定性、以及企业规模成形需要一定时间有关。

以下为所有字段的相关性热力图，如图 4.13：

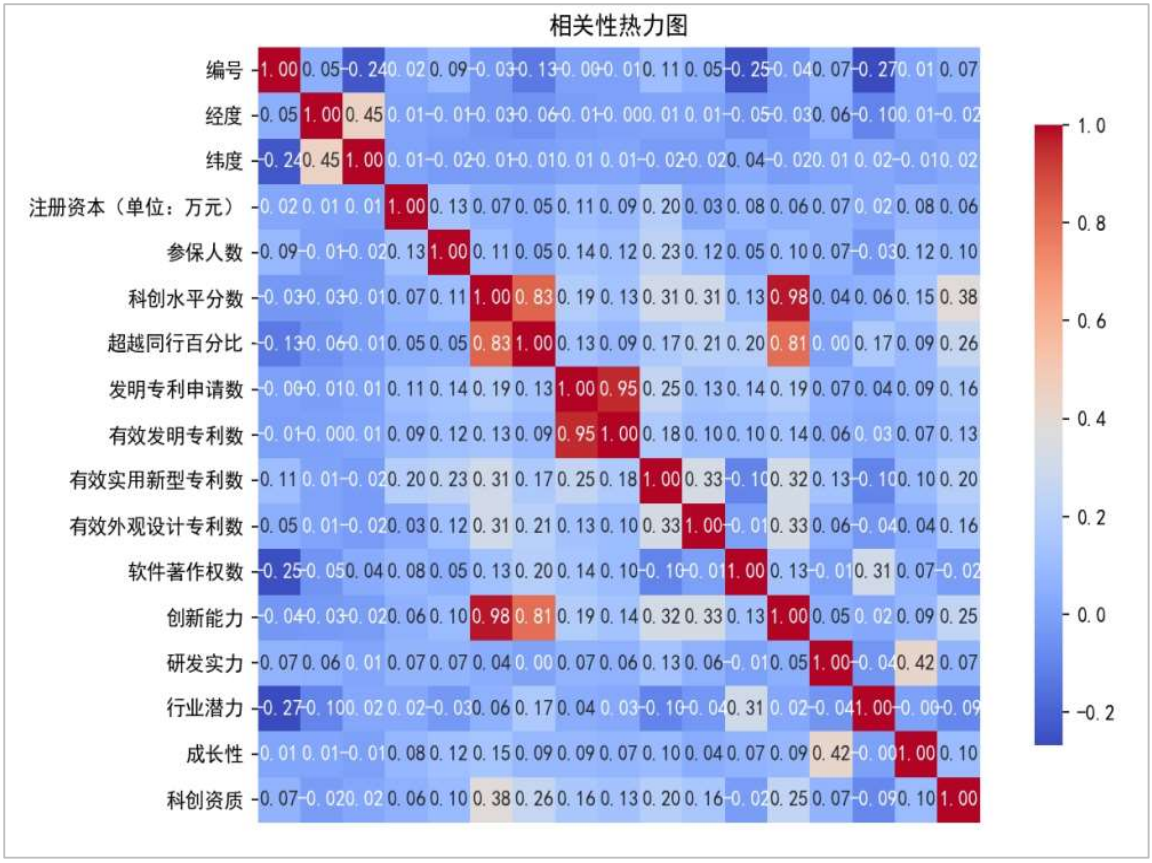

图 4.13 所有字段相关性热力图

从热力图中可以看出，科创水平分数与以下字段有较高的正相关性：超越同行百分比（相关性系数：0.83）、创新能力（相关性系数：0.98）、研发实力（相关性系数：0.51）、有效发明专利数（相关性系数：0.19）、有效实用新型专利数（相关性系数：0.31）。

科创水平分数的高低主要受创新能力的驱动，并且与企业在行业中的竞争优势（超越同行百分比）密切相关。同时，研发实力和专利数量（尤其是发明专利）也是重要的辅助因素。体现出企业需要在提升核心创新能力的同时，加

大研发投入，注重高质量专利的申请和布局，才能在科创水平上实现突破性提升。

4.1.2 数据规律与趋势发现

通过对“专精特新中小企业”的基础地段可视化的分析与整体相关性计算，可以发现这些企业在规模、行业分布、区域集中度和科技创新能力等方面呈现出显著特征：

1. 在企业规模上，大型企业占据主导地位，但也包含一定比例的中小企业，体现了认定标准对企业市场影响力和科技创新能力的双重要求，同时兼顾不同规模企业的全面发展需求。

2. 从行业分布来看，企业主要集中在高科技和创新驱动领域，如研究和试验发展、计算机通信设备制造、电气机械制造等，反映了这些行业在推动技术进步和产业升级中的关键作用，尤其在数字化、智能化和绿色发展领域展现出强劲的增长趋势。

3. 区域分布方面，企业主要集中在广东省珠三角地区的核心城市，如东莞、广州和佛山，这些区域不仅企业数量多且等级高，表现出明显的产业集聚效应和高质量发展优势。这一分布得益于珠三角地区完善的产业链、高效的政策支持以及优越的创新环境，而粤东和粤西地区则由于基础设施和政策支持相对不足，企业分布较为分散。

4. 科创水平的高低与企业的创新能力、超越同行的竞争优势、研发实力及专利数量密切相关，高质量专利的积累和创新投入成为提升企业科创水平的重要因素。

综合来看，“专精特新中小企业”的发展与政策支持、市场需求及技术进步密切相连，为推动区域经济和产业升级提供了重要动力，也为政策优化和资源分配指明了方向。

4.2 关联分析

本章主要围绕企业科创水平的影响因素展开关联分析，探讨不同指标与企业科创水平之间的关系。通过引入多维度的变量，包括科创能力、知识产权、

企业认定类型以及所在区域等，系统性地分析各因素对科创水平的影响。具体而言，在下文分析企业研发投入、创新成果、专利申请与授权情况、参保人数、企业多重认定类型以及不同区域分布对企业科创平均评分的推动作用。通过数据可视化和逐步深入的分析，本章旨在揭示影响企业科创水平的关键因素，并为优化企业创新能力、提升区域科创水平提供数据支持与决策参考。

4.2.1 变量关系分析

探究科创指标字段分别为创新能力、研发实力、行业潜力、成长性、科创资质与科创水平的关系，如图 4.14：

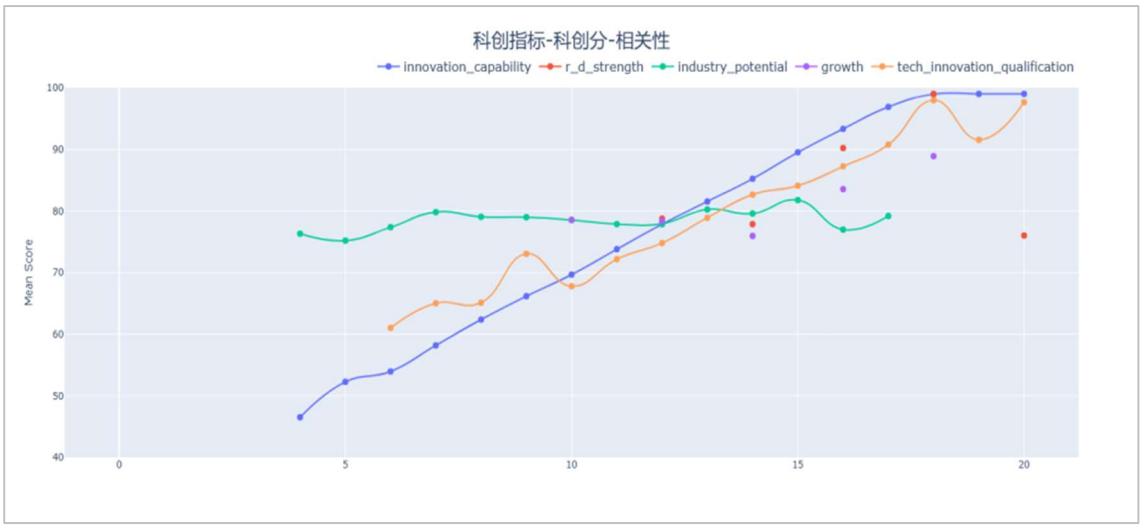

图 4.14 科创指标与科创水平关系图

主要为创新能力、行业潜力、科创资质这三项呈现出稳定的线性关系，而研发实力与成长性由于较多数据缺失未能体现。结合相关性热力图的结果，对创新能力与科创资质着重分析：创新能力与科创水平呈现出显著的线性正相关，是提升科创水平的核心驱动因素，表明企业在提升核心竞争力的过程中，应优先聚焦技术研发、产品创新和技术成果转化等环节，以最大化创新资源的利用效率。科创资质也表现出较高的相关性，但在部分区间出现波动，反映出企业在特定发展阶段可能面临资源分配不足或边际收益递减等问题，亟需优化研发投入策略，确保创新能力的可持续性。

探究发明专利申请数、有效发明专利数、有效实用新型专利数、有效外观设计专利数、软件著作权数与科创水平的关系，如图 4.15：

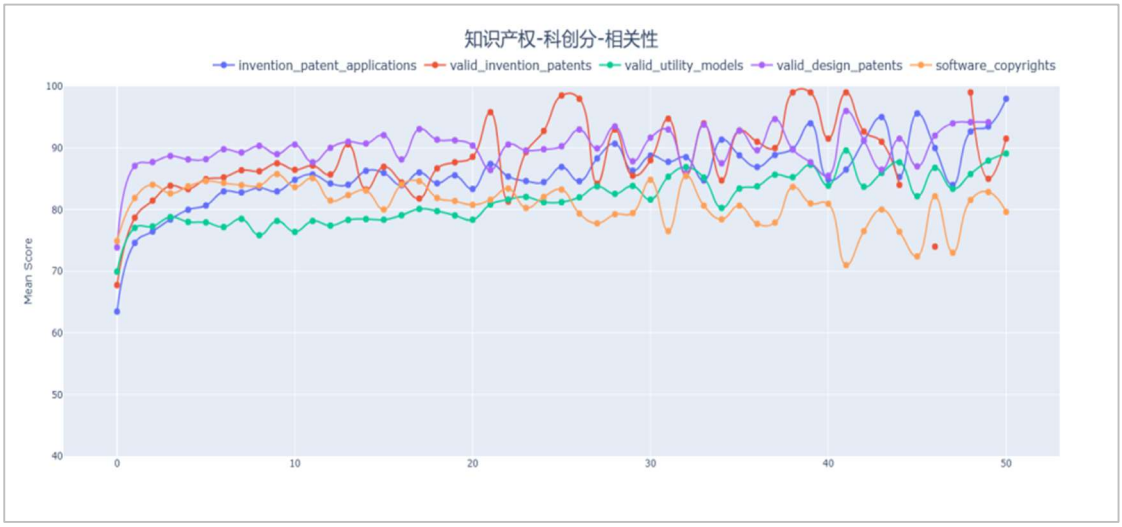

图 4.15 知识产权与科创水平关系图

从图表中可以看出，知识产权相关指标与科创分数之间存在一定的关联性，但各指标对科创水平分数的影响程度和趋势有所不同。其中，有效发明专利数与科创分数的相关性最为显著，其曲线整体稳定并呈现较高的波动幅度，体现发明专利的数量和质量是衡量企业科技创新能力的重要标准。同时发明专利申请数也与科创分数高度相关，体现了企业在知识产权积累初期的研发投入对于科创水平的积极促进作用。

有效实用新型专利数和有效外观设计专利数的波动较小，对科创分数的贡献相对平稳。这表明，这两类专利在一定程度上补充了企业的知识产权布局，但其影响力不及发明专利。此外，软件著作权数量在部分区间对科创分数的影响较为显著，尤其是在信息技术和软件行业，其曲线走势与科创分数呈现一定的协同性，反映了数字化与智能化发展趋势对企业创新能力的推动作用。

深入分析参保人数与科创水平等级的关系，如图 4.16：

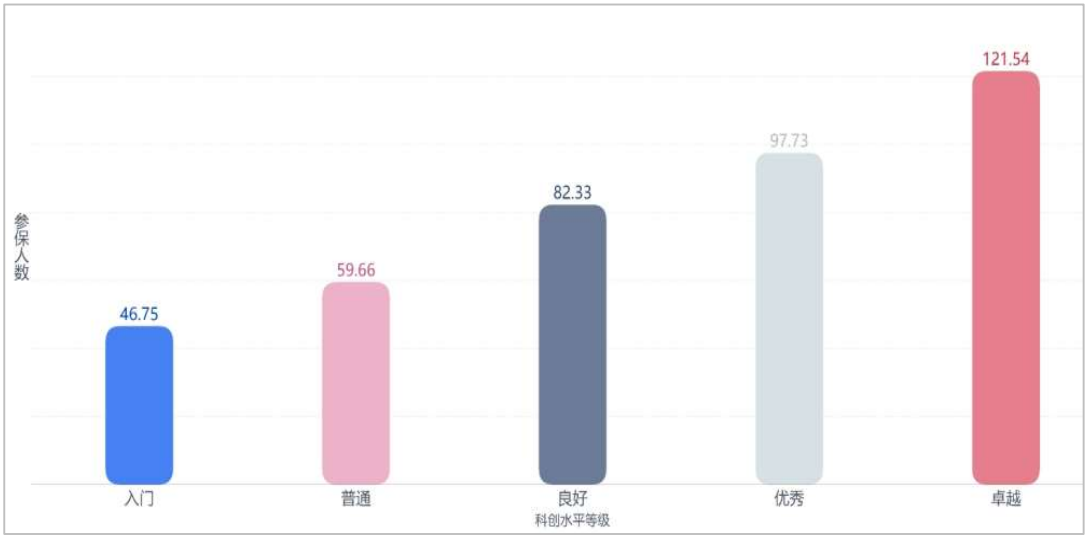

图 4.16 参保人数与科创水平等级关系图

从图中可以看出，参保人数与科创水平等级之间存在显著的正相关关系。随着科创水平从“入门”到“卓越”逐步提高，参保人数也呈现出明显的增长趋势。体现出企业的科创能力与其员工规模之间有较强的关联性，科创能力更高的企业通常具备更大的规模，表明企业在跨越高科创门槛时需要投入更多的资源和吸纳更多的人才支持

发现企业有多个认定，以出现频率最高的六个认定进行组合，即最后一个类别为同时具有所有认定的企业，对其科创水平进性分析，如图 4.17：

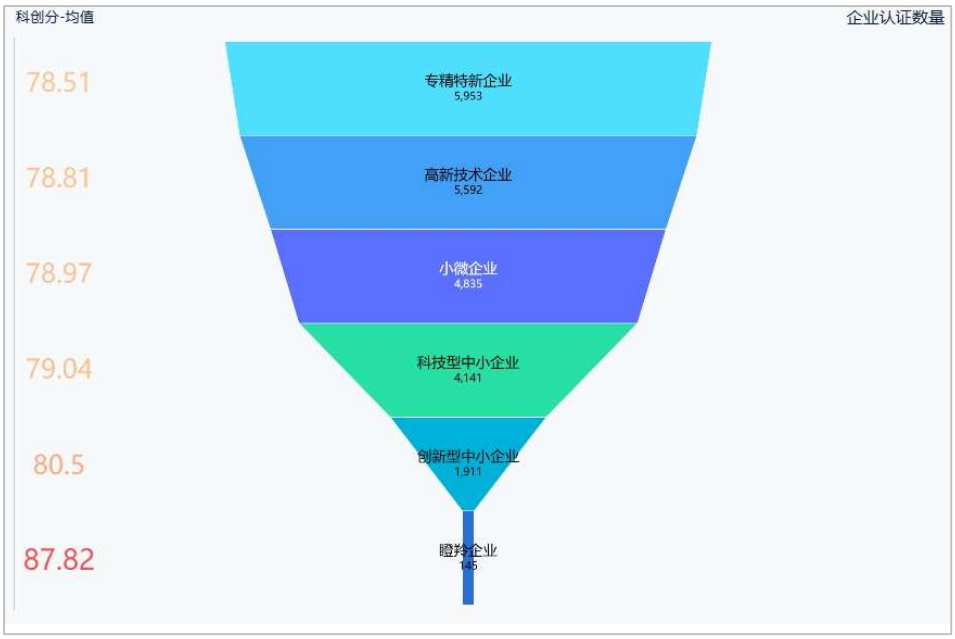

图 4.17 认定数量与科创水平关系图

可以看出拥有更多类型认定的企业，其科创水平更高。企业被认定类型的增加，平均科创评分呈现出持续上升的趋势。例如，仅属于“专精特新中小企业”的企业平均科创评分为 78.81，而当企业同时拥有“专精特新中小企业”“高新技术企业”“小微企业”等多重认定时，平均评分达到 78.97 至 80.50 不等。当企业获得包含“瞪羚企业”在内的所有类型认定时，其平均科创评分显著上升至 87.82。这表明，多重认定能够有效反映企业在创新能力、技术研发和市场竞争力等方面的综合优势，认定类型越多，企业的整体科创水平越高。

但在整体企业群体中，能够获得多重认定的企业具有较高的稀缺性。这种稀缺性不仅反映了企业在创新资源、技术储备和成长潜力上的核心竞争力，也揭示出高水平科创企业在认定门槛上的严格性。尤其是包含“创新型中小企业”和“瞪羚企业”的组合，尽管企业数量大幅减少，但它们显著提高了科创水平的整体评分，说明高成长性和创新驱动型企业是推动科技创新发展的重要力量。因此，未来政策应鼓励企业争取更多类型的认定，通过资源倾斜和激励机制，引导更多企业向多重认定方向发展，进而全面提升整体科创水平与经济活力。

最后分析不同区域的科创水平，如图 4.18：

图 4.18 企业所属区域与科创水平关系图

从图表中可以看出，不同区域的科创水平存在一定差异，各区域平均科创评分波动在 72.9 至 82.13 之间。整体而言，广州的多个区域表现出较高的科创

水平，例如广州天河区（82.13）和广州海珠区（81.55），显示了这些区域在科技创新资源、企业发展和政策扶持上的优势。与之相比，部分区域的科创水平相对较低，例如广州增城区（72.9）和广州南沙区（74.21），可能受到产业基础薄弱或创新资源不足的影响。

值得注意的是，中山市和珠海市的科创水平也达到一定高度，与广州核心城区差距较小，这表明这些地区可能在区域协同创新政策或特色产业发展上取得了积极成效。然而，广州市内不同城区之间的差异显示了资源和政策分布的非均衡性。未来可以通过推动区域间资源整合和创新协同发展，进一步提升整体区域的科创能力，缩小低水平区域与高水平区域之间的差距，从而实现全区域范围内的创新驱动均衡发展。

4.2.2 潜在关联规律揭示

通过对专精特新中小企业科创水平的多维度关联分析，发现企业科创水平受到多方面因素的显著影响。首先，企业的创新能力、科创资质与科创水平呈现较强的正相关性。这表明在提升科创能力的过程中，加强企业创新资源的投入和优化其成长路径，是提高科创水平的关键。同时，多种技术创新资质的认定对企业科创水平也有显著的提升作用，企业获得更多资质认证往往意味着更高的综合创新能力与市场竞争力。

知识产权是企业科创水平的重要支撑要素，其中发明专利申请数、有效专利数量和软件著作权数量等指标在一定程度上反映了企业的创新成果与技术积累。分析表明，知识产权的积累与科创水平成正向关联，尤其是高价值的知识产权，如有效发明专利，对科创水平的提升尤为显著。同时，参保人数作为企业规模和活跃度的间接指标，也与科创水平呈正相关关系，表明员工规模较大的企业往往在创新活动中具有更强的资源支持和市场影响力。

从区域分布来看，不同区域的企业在科创水平上表现出显著差异，广州核心城区整体高于其他区域，而部分外围区域的科创水平略低。区域差异不仅反映了创新资源的集中程度，也与区域经济发展水平和政策扶持力度密切相关。因此，进一步优化区域创新资源分配，推动外围区域企业提升科创能力，将有助于整体科创水平的均衡发展。总体而言，本章揭示了企业科创水平与多项指

标间的潜在关联规律，为后续政策制定和企业创新发展方向提供了重要的参考依据。

5 模型构建与预测

5.1 特征工程

在数据分析与建模过程中，特征工程是提升模型效果的关键步骤。通过科学合理的特征选择与提取，可以有效降低噪声、提升模型预测能力。本节主要描述特征工程的具体实施方法和取得的效果。

5.1.1 特征选择与提取

在数据处理阶段，剔除了与科创水平预测无直接关联或难以客观测量的字段，包括以下几类数据：

1. 无关特征：如企业的地理位置（longitude 和 latitude）、企业地址（address）、成立时间（founding_date）和企业名称（company_name），这些字段对预测科创水平没有实际意义。

2. 难以客观提供的特征：如企业的超越同行比例（exceed_percent）和科创水平指标（如 innovation_capability、r_d_strength、industry_potential 等），这些数据虽然在分析中可能有效，但实际使用中因主观性或数据获取难度较高，不利于应用推广。

3. 冗余信息：如公司注册资本（capital）和业务范围（business_scope），它们在当前预测任务中意义有限。

同时，对城市分区（city_district）、企业类型（type）、所属行业（industry_lv2）采用独热编码（One-Hot Encoding）进行处理，转化为模型可以识别的数值特征，从而保留类别信息的同时避免顺序误导。

主要代码如下：

```
# 无关的的数据
drop_columns = ['id', 'address', 'longitude', 'latitude', 'founding_date',
'company_name', 'capital',
                'matched_certification_list',
'business_scope','innovation_capability', 'r_d_strength',
                'industry_potential', 'growth','tech_innovation_qualification',
'exceed_percent', 'grade', 'scale']
df = df.drop(drop_columns, axis=1)


# 独热编码处理
df = pd.get_dummies(df, columns=['city_district', 'type', 'industry_lv2'],
dtype='uint8')
```

通过上述特征选择与提取，最终保留了与企业科创水平密切相关的特征，为模型提供了有效的数据基础。

5.1.2 特征工程目的与效果

特征工程的主要目的是通过剔除无关、冗余或难以量化的特征，简化模型输入，提高模型的可用性和实际预测能力。此外，特征工程能帮助后续使用者轻松采集与填写特征数据，降低预测模型在应用中的门槛。

下表展示了最终保留特征的数据类型，表明经过特征工程后的数据具备较好的可用性和结构性，如表 5.1：

表 4.1 特征工程后字段类型表

字段名称	数据类型	说明
capital_cny_w	float64	企业注册资本（单位：万元）
insured_number	int64	企业参保人数
Score	int64	科创评分
字段名称	**数据类型**	**说明**

invention_patent_applications	int64	发明专利申请数量
valid_invention_patents	int64	有效发明专利数量
valid_utility_models	int64	有效实用新型专利数量
valid_design_patents	int64	有效外观设计专利数量
software_copyrights	int64	软件著作权数量
city_district	uint8	城市分区（独热编码后字段）
type	uint8	企业类型（独热编码后字段）
industry_lv2	uint8	行业二级分类（独热编码后字段）

通过上述步骤，特征工程有效提高了模型的学习效率，为后续的建模与预测提供了坚实基础，同时显著提升了模型在实际应用中的适配性。

5.2 模型训练与评估

5.2.1 模型选择与训练

在本研究中，尝试了多种模型来完成科创评分的回归预测任务，包括线性回归、岭回归、套索回归、弹性网回归、决策树回归、随机森林回归和梯度提升树回归。模型选择的依据是其对数据非线性关系的建模能力以及预测精度。最终，梯度提升树回归（Gradient Boosting Regressor）由于其较低的误差（MSE）和较高的 R^2 值，表现最佳，被选为最终预测模型。

主要代码如下：

```
# 定义自变量和因变量
X = df.drop(columns=['score'])
y = df['score']

# 拆分数据集为训练集和测试集
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

```
# 线性回归模型（需要添加常数项）
X_train_with_const = sm.add_constant(X_train)
X_test_with_const = sm.add_constant(X_test)
linear_model = sm.OLS(y_train, X_train_with_const).fit()
print("线性回归 MSE:", mean_squared_error(y_test,
linear_model.predict(X_test_with_const)))

# 岭回归模型（不需要添加常数项）
ridge_model = Ridge(alpha=1.0)
ridge_model.fit(X_train, y_train)
ridge_preds = ridge_model.predict(X_test)
print("岭回归 MSE:", mean_squared_error(y_test, ridge_preds))

# 套索回归模型（不需要添加常数项）
lasso_model = Lasso(alpha=0.1)
lasso_model.fit(X_train, y_train)
lasso_preds = lasso_model.predict(X_test)
print("套索回归 MSE:", mean_squared_error(y_test, lasso_preds))

# 弹性网回归模型（不需要添加常数项）
elastic_net_model = ElasticNet(alpha=0.1, l1_ratio=0.5)
elastic_net_model.fit(X_train, y_train)
elastic_net_preds = elastic_net_model.predict(X_test)
print("弹性网回归 MSE:", mean_squared_error(y_test, elastic_net_preds))

# 决策树回归模型（不需要添加常数项）
tree_model = DecisionTreeRegressor(random_state=42)
tree_model.fit(X_train, y_train)
```

```
tree_preds = tree_model.predict(X_test)
print("决策树回归 MSE:", mean_squared_error(y_test, tree_preds))


# 随机森林回归模型（不需要添加常数项）
forest_model = RandomForestRegressor(n_estimators=100, random_state=42)
forest_model.fit(X_train, y_train)
forest_preds = forest_model.predict(X_test)
print("随机森林回归 MSE:", mean_squared_error(y_test, forest_preds))


# 梯度提升树模型（不需要添加常数项）
gbt_model = GradientBoostingRegressor(n_estimators=100, learning_rate=0.1,
random_state=42)
gbt_model.fit(X_train, y_train)
gbt_preds = gbt_model.predict(X_test)
print("梯度提升树回归 MSE:", mean_squared_error(y_test, gbt_preds))
```

5.2.2 模型性能评估与优化

在进行模型训练后，采用以下评估指标对模型性能进行综合评价：

1. 均方误差（Mean Squared Error, MSE）

含义与作用：均方误差是衡量模型预测值与实际值差异的常用指标，它反映了模型预测的准确性。MSE 值越低，说明模型的预测精度越高。

计算方式：MSE 是预测值与实际值之差的平方的平均值。

计算公式为：

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

其中，（ yi ）是第（ i ）个实际值，（ $\hat{y}_i$）是第（ i ）个预测值，（ n ）是样本数量。

2. 均方根误差（Root Mean Squared Error, RMSE）

含义与作用：均方根误差是 MSE 的平方根，它提供了预测值与实际值之间差异的标准度量，单位与原数据单位相同，便于直观理解预测误差的大小。

计算方式：RMSE 是 MSE 的平方根，

计算公式为

$$[RMSE = \sqrt{MSE}]$$

3. 平均绝对误差（Mean Absolute Error, MAE）

含义与作用：平均绝对误差表示预测值与实际值之间差异的平均绝对值，它对异常值不如 MSE 敏感，因此可以提供模型稳健性的额外信息。

计算方式：MAE 是预测值与实际值之差的绝对值的平均数。

计算公式为：

$$[MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|]$$

4. 决定系数（Coefficient of Determination, R^2）

含义与作用：决定系数也称为拟合优度，它表示模型对数据变异的解释程度。R^2 的值介于 0 到 1 之间，越接近 1 表明模型解释力越强。

计算方式：R^2 计算的是实际值与均值之差的总平方和与实际值与预测值之差的总平方和的比值。

计算公式为：

$$[R^2 = 1 - \frac{SS_{res}}{SS_{tot}}]$$

其中，（ SSres ）是残差平方和，（ SStot ）是总平方和。

通过上述指标的综合评价，得以全面了解模型的性能，并针对不足进行优化。从结果来看，线性回归等简单模型因无法捕捉复杂特征关系，表现出较高的误差。而决策树和随机森林虽然能显著降低误差，但梯度提升树回归在所有

指标上表现最佳，具有更高的预测能力。通过超参数调优（如调整迭代次数、收缩率）进一步提高了模型性能，最终在交叉验证中实现了稳定的误差控制。

5.3 模型选择与预测

5.3.1 模型性能比较

不同模型在测试集上的性能如下，如图 5.1：

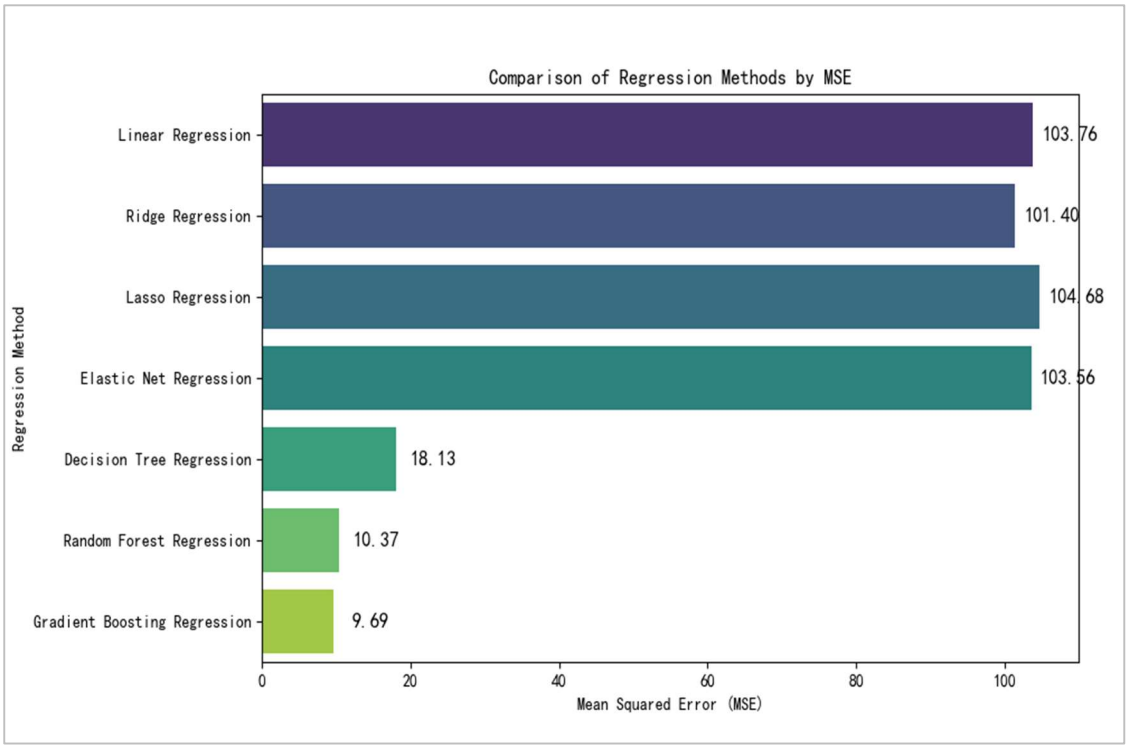

图 5.1 模型 MSE 对比图

在本研究中，对多种回归模型在数据集上的性能进行了比较。结果显示，简单模型如线性回归、岭回归和 Lasso 回归在预测精度上表现较差，其均方误差（MSE）较高，这表明它们无法有效捕捉数据的复杂特征关系。相比之下，决策树回归、随机森林回归和梯度提升树回归等非线性模型在降低误差方面表现更为出色，其中梯度提升树回归的 MSE 值最低，仅为 9.69，显著优于其他模型，显示出其强大的预测能力和对复杂数据关系的建模能力。

进一步分析表明，集成学习方法如随机森林和梯度提升树通过组合多个决策树，有效提高了预测的准确性和稳定性。特别是梯度提升树回归，通过连续

构建和组合弱学习器，不仅降低了预测误差，而且在处理数据中的非线性关系方面表现突出，使其成为本研究中最合适的预测模型。

5.3.2 预测结果评估

最终模型采用梯度提升树回归模型，其预测结果与实际值的表现如下：MSE 为 9.69，RMSE 为 3.11，MAE 为 2.41，决定系数 R^2 达到 0.917。基于误差指标可以看出，该模型的均方误差和均方根误差较小，表明其对目标变量的预测误差幅度较低。同时，MAE 的结果显示了预测误差的稳定性，模型对数据中多数样本的预测表现良好。而决定系数接近 1，则表明模型对目标变量方差的解释能力较强，整体拟合效果优异。

此外，通过交叉验证进一步验证了模型的稳健性和泛化能力。在交叉验证过程中，模型的 MSE 为 9.35，RMSE 为 3.06，两个指标均与原始模型训练的结果接近，说明模型在不同样本集上的预测误差和误差幅度保持一致。这种一致性证明了模型具有较高的可靠性，不易过拟合或受数据集分割的影响。

因此梯度提升树回归模型不仅在训练数据上的拟合效果优异，而且在验证数据集上的表现也十分稳定，展现出强大的预测能力。其低误差和高 R^2 使其成为科创评分回归预测任务的理想选择，为后续的企业评估和决策提供了可靠依据。

为了更直观地了解梯度提升树回归模型的效果，对模型的各项结果进行可视化分析，包括特征重要性、相对误差分布、残差分析、预测值与真实值的可视化，如图 5.2：

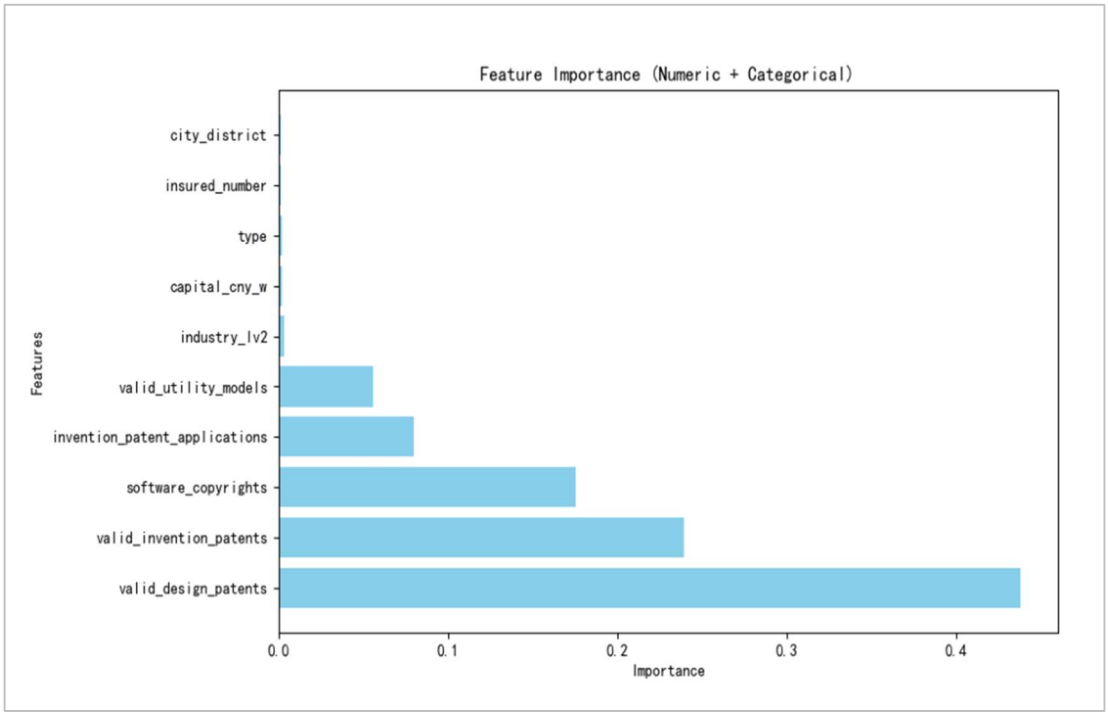

图 5.2 特征重要性可视化图

从特征重要性分析结果可以看出，模型对科创水平的预测高度依赖于企业技术能力和创新成果相关的特征，尤其是有效外观专利数、有效发明专利数和软件著作权，这些变量的重要性分数显著高于其他特征，表明它们对企业的创新能力和科创评分具有决定性作用。同时发明专利申请数和有效实用新型专利数等变量也表现出一定的贡献，进一步反映了企业技术开发与保护活动的重要性。相较之下，与企业规模、地区或类型相关的变量（如注册资本、参保人数、企业类型和所在区域）对模型的贡献较低，表明这些特征对科创评分的解释能力较弱。这一结果凸显了创新成果在科创评价体系中的核心地位，一定程度上体现了所属区域、行业类型等方面并不能体现出企业的科创水平，只能代表有一定的基础能力，而关键在于其创新成果。

对相对误差分布进行可视化分析，如图 5.3：

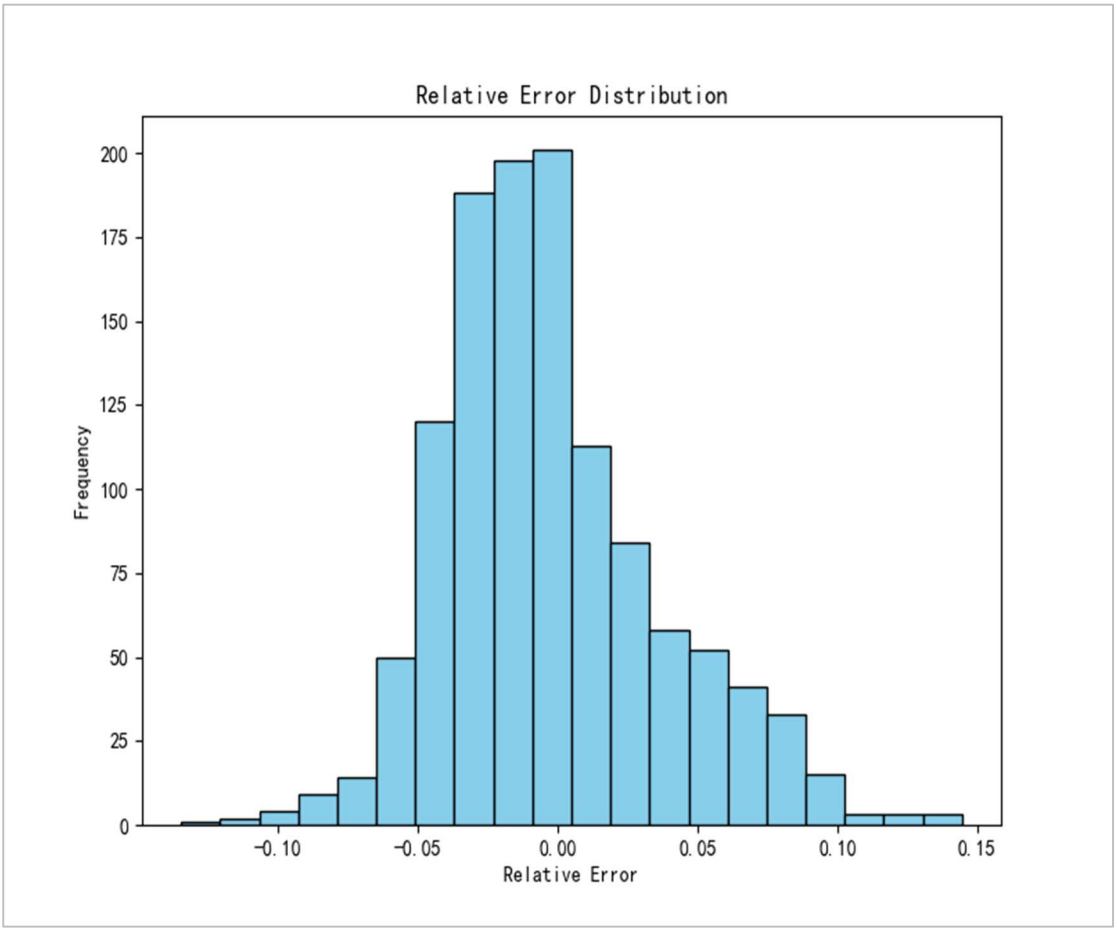

图 5.3 相对误差分布图

该图展示了模型预测的相对误差分布情况。从分布来看，误差呈现出明显的正态分布特性，大部分预测值的相对误差集中在 0 附近，误差范围主要分布在 -0.05 到 0.05 之间，这表明模型具有较高的预测精度。此外，极端误差（大于 ±0.1 的部分）较少，进一步表明模型在整体上对数据具有较好的拟合能力和稳定性。

误差分布的对称性和集中性说明模型对目标变量的解释力较强，预测结果与实际值之间的偏差较小，未出现显著的系统性误差，这反映了梯度提升树模型的可靠性和稳健性。

对残差进行可视化分析，如图 5.4：

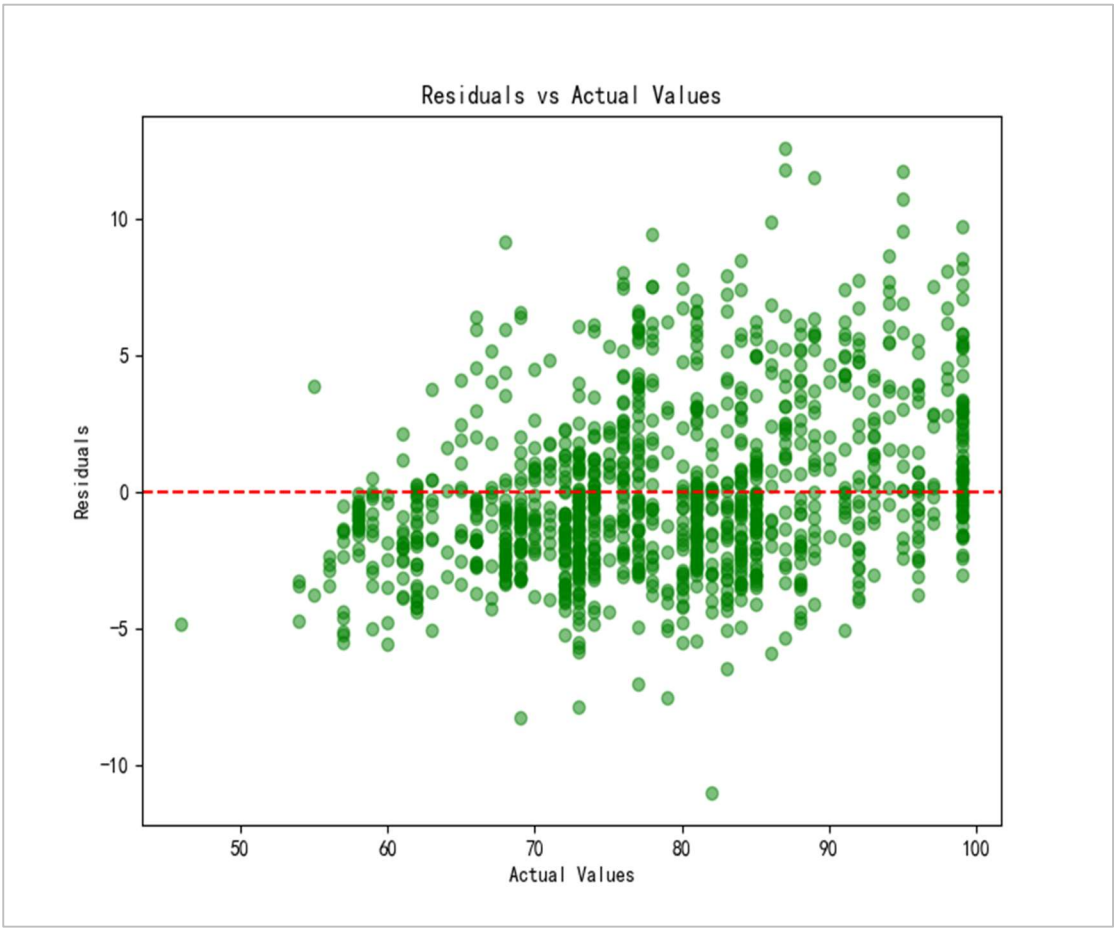

图 5.4 残差可视化图

残差与实际值散点图显示，大部分残差点分布在零附近，且在不同实际值范围内无明显偏差，表明模型整体拟合效果较为均匀，未出现显著过拟合或欠拟合现象。残差在正负方向分布较为平衡，但在部分高实际值区间离散程度略大，可能表明模型在该部分数据的拟合效果稍逊。总体上模型能够较好捕捉数据特征，预测误差分布合理，具备较强的稳健性和泛化能力。

对预测值与真实值进行可视化分析，如图 5.5：

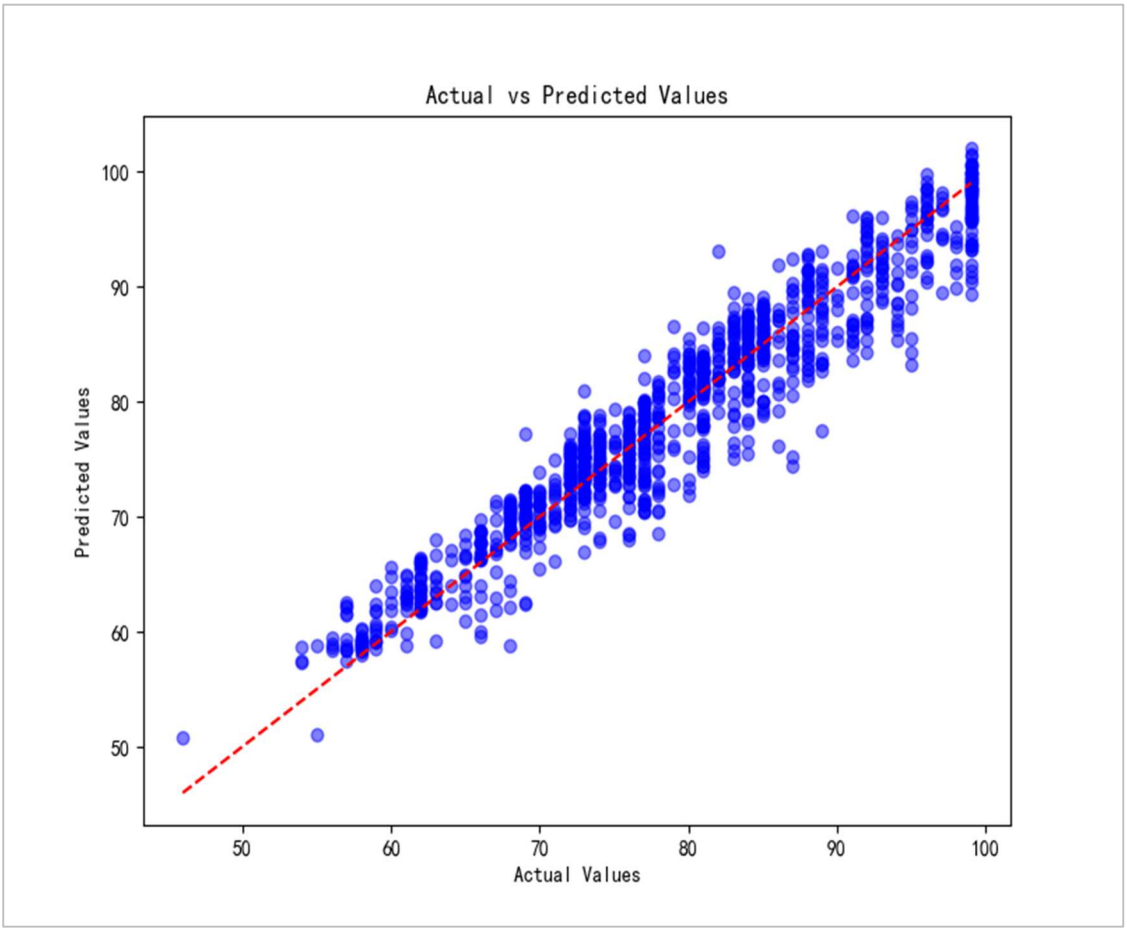

图 5.5 预测值与真实值对比图

预测值与实际值的散点图显示，数据点大体沿对角线分布，表明模型预测值与实际值之间具有很高的吻合度。红色虚线为理想拟合线（即预测值等于实际值的情况），大部分点均较为接近该拟合线，表明模型具有较高的准确性。在较高和较低实际值区间，数据点的偏离程度略微增大，反映了这些区间内数据的复杂性或样本数量较少对模型训练的影响。该图表明模型在大多数范围内表现良好，能够较为精准地进行评分预测。

从上述分析可以看出，通过对模型的预测结果进行多维度评估，包括误差指标和多种可视化分析，进一步验证了梯度提升树模型在科创评分回归任务中的优越表现。从误差指标来看，模型的 MSE、RMSE、MAE 较低，R^2 接近 1，表明模型具有良好的拟合能力和预测精度。可视化结果进一步支持了这一结论：特征重要性分析展示了模型对各指标的依赖程度，残差分析和相对误差分布表明模型预测的偏差较小且分布合理，而预测值与实际值的对比验证了模型在大部

分区间的预测准确性。整体上该模型不仅能够较好地捕捉数据中的特征关系，还具有较高的稳定性与鲁棒性，适用于该领域的实际应用。

5.4 web 应用

在当今的数据科学领域，将复杂的机器学习模型转化为易于使用的 web 应用程序是一个重要的趋势。Streamlit 是一个新兴的 Python 库，它为数据科学家和机器学习工程师提供了一种高效的方式来构建和部署数据驱动的 web 应用。本节将详细介绍 Streamlit 库，并将其运用在该项目模型预测中。

5.4.1 streamlib 介绍

Streamlit 的优势在于其快速原型设计和无需复杂的配置。与 Django 和 Flask 相比，Streamlit 更加专注于数据科学和机器学习项目的快速部署，并且简化了 web 应用的开发流程，让开发者可以不依赖前端知识就能构建交互式应用。而 Django 和 Flask 则是全栈 web 框架，它们提供了更全面的 web 开发工具和组件，适合构建大型、复杂的 web 应用。虽然 Django 和 Flask 提供了更细粒度的控制和更丰富的功能，但 Streamlit 更适合需要快速将模型转化为可交互 web 应用的情况下。

5.4.2 主要代码及作用

编程环境为：windows10、python3.9、pycharm2022

主要库版本为：pandas==2.2.2、numpy==2.1.2、joblib==1.4.2、scikit-learn==1.5.2、streamlit==1.39.0

主要代码及注释：

```
# 设置页面的标题、图标和布局
st.set_page_config(
    page_title="企业科创水平预测器",  # 页面标题
    page_icon=":bar_chart:",  # 页面图标
    layout='wide',
)
```

```
# 使用侧边栏实现多页面效果
with st.sidebar:
    st.image('数据分析/项目五-pro/logo.png', width=300)
    st.title('请选择页面')
    page = st.selectbox("请选择页面", ["简介页面", "预测分类页面"],
label_visibility='collapsed')

if page == "简介页面":
    st.title("企业科创水平预测器 :bar_chart:")
    st.header('模型介绍')
    st.markdown("""模型用于预测公司科创指标评分。训练集中包含了关于公司的一些特征，
如注册资本、专利数量、软件著作权数量等。""")
    st.header('特征重要性')
    st.image('数据分析/项目五-pro/特征重要性.png')
    st.header('相对误差分布')
    st.image('数据分析/项目五-pro/相对误差分布.png')
    st.header('预测值与真实值的可视化')
    st.image('数据分析/项目五-pro/预测值与真实值的可视化.png')

elif page == "预测分类页面":
    st.header("预测公司评分")
    st.markdown("这个 Web 应用是基于公司数据集构建的模型。只需要输入一些信息，就可以
预测公司的评分，使用下面的表单开始预测吧！")

    # 运用表单和表单提交按钮
    with st.form('user_inputs'):
        capital_cny_w = st.number_input('注册资本（万元）', min_value=0.0)
        insured_number = st.number_input('被保险人数', min_value=0)
        invention_patent_applications = st.number_input('发明专利申请数',
min_value=0)
        valid_invention_patents = st.number_input('有效发明专利数', min_value=0)
        valid_utility_models = st.number_input('有效实用新型专利数', min_value=0)
        valid_design_patents = st.number_input('有效外观设计专利数', min_value=0)
        software_copyrights = st.number_input('软件著作权数', min_value=0)

        # 分类变量选择
        city_district = st.selectbox('选择公司所在城市区', options=['东莞市-', '广
```

```
州市黄埔区', '广州市天河区', '中山市-', '广州市番禺区'])
        company_type = st.selectbox('选择公司类型', options=['有限责任公司(自然人投资或控股)', '其他有限责任公司', '有限责任公司(自然人独资)', '有限责任公司（非自然人投资或控股的法人独资）', '有限责任公司(法人独资)'])
        industry_lv2 = st.selectbox('选择公司行业', options=['研究和试验发展', '计算机、通信和其他电子设备制造业', '电气机械和器材制造业', '软件和信息技术服务业', '金属制品业'])

        submitted = st.form_submit_button('预测分类')

    if submitted:
        # 将用户输入数据构建为字典
        input_data = {
            'capital_cny_w': capital_cny_w,
            'insured_number': insured_number,
            'invention_patent_applications': invention_patent_applications,
            'valid_invention_patents': valid_invention_patents,
            'valid_utility_models': valid_utility_models,
            'valid_design_patents': valid_design_patents,
            'software_copyrights': software_copyrights,
            'city_district': city_district,
            'type': company_type,
            'industry_lv2': industry_lv2
        }

        # 转换为 DataFrame
        df_input = pd.DataFrame([input_data])

        # 对输入数据的分类变量进行独热编码
        df_input = pd.get_dummies(df_input, columns=['city_district', 'type', 'industry_lv2'], dtype='uint8')

        # 确保输入数据列顺序和训练数据一致
        for col in columns:
            if col not in df_input.columns:
                df_input[col] = 0  # 添加缺失的列

        df_input = df_input[columns]  # 按照训练数据的列顺序排列
```

```
    # 使用模型进行预测
    prediction = model.predict(df_input)
    result = int(prediction[0])

    st.write(f'根据您输入的数据，预测该公司的评分是：**{result}**')
```

该代码段创建了一个基于 Streamlit 的 web 应用，用于展示企业科创水平预测器。首先设置了页面的标题、图标和布局，并通过侧边栏实现了多页面导航。在"简介页面"中，展示了模型介绍、特征重要性图表、相对误差分布图以及预测值与真实值的关系图。在"预测分类页面"中，用户可以通过填写表单输入公司信息，提交后，应用将使用机器学习模型预测公司的科创评分，并将结果展示给用户。整个应用流程简化了用户获取预测结果的步骤，提高了模型的实用性和可访问性。

5.4.3 应用效果

介绍页，如图 5.6、5.7：

图 5.6 模型预测介绍页 1

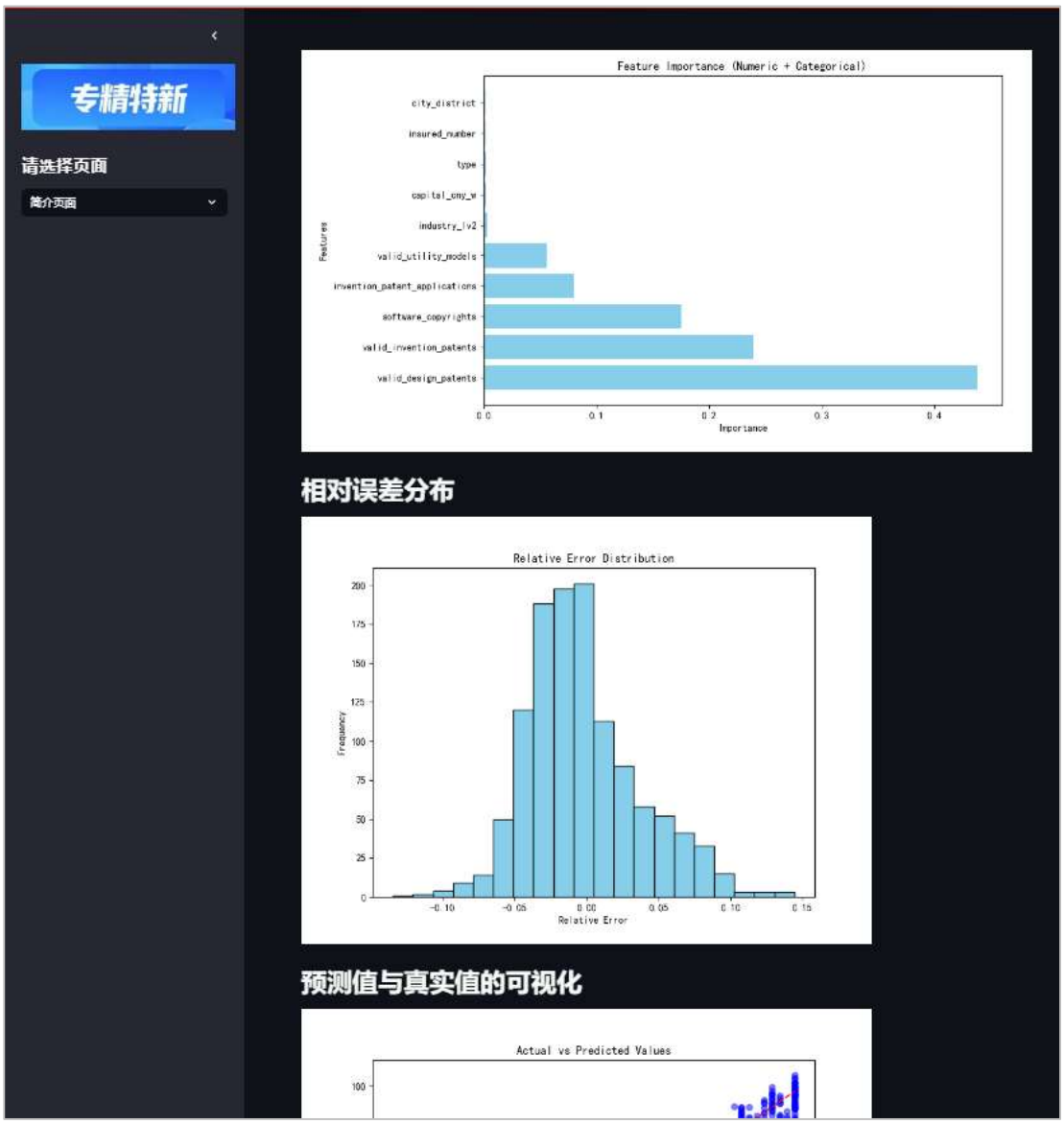

图 5.7 模型预测介绍页 2

使用者可以在该页面中对模型有一定程度的了解，并通过特征重要性图认识到影响科创水平的关键指标。

预测页面，如图 5.8、5.9、5.10：

图 5.8 模型预测页 1

图 5.9 模型预测页 2

图 5.10 模型预测页 3

使用者可以在该页码进行企业数据的输入，其中分类变量采取选择框的形式，但数据中分类变量选项过多容易造成使用者反感、且考虑到分类变量对结果的影响程度有限，故仅提供频率最高的前五项供使用者进行选择。

6 总结与展望

6.1 总结

6.1.1 研究成果概述

本研究针对“专精特新”中小企业的行业特点及评级需求，进行了数据分析与 Web 预测系统的开发工作。研究基于真实企业数据，通过数据分析和挖掘，揭示了企业创新能力、专利数量、企业规模等特征与评分之间的潜在关系。具体而言，利用 Python 可视化库（如 Matplotlib、Seaborn）和商业智能工具 FineBI，完成了大量数据可视化任务，对行业内企业的分布特征、指标相关性以及企业创新水平的区域差异进行了系统性分析，为相关政策制定和企业优化提供了数据支持。

在预测系统开发方面，通过构建梯度提升树（GBDT）模型对企业评分进行精准预测。模型在训练阶段表现出优异的性能，预测误差低且稳健性强，主要指标（如 MSE、RMSE 和 R^2）均达到较高水准。为了确保模型的实际应用价值，还进行了交叉验证，验证结果进一步支持模型的可靠性。

此外，研究采用 Streamlit 框架实现了一个交互式的 Web 预测系统，用户可以通过简单的界面操作输入企业特征数据，实时获得评分预测结果。系统同时集成了模型展示功能，如特征重要性可视化、预测误差分析等，为用户提供了直观的辅助决策工具。结合 PyCharm 和 Python 3 等开发环境，整个研究从数据预处理、特征工程到模型训练与部署均完成了高效的集成开发，最终形成了一套以数据分析为核心、预测功能为导向的研究。

6.1.2 创新点与不足

创新点：

1. 提出基于企业行业特性和专精特新背景的特征选择方案，并通过梯度提升树模型实现高精度评分预测。

2. 融合了静态与交互式数据可视化技术，既满足学术研究需求，也为企业实践提供了直观的辅助工具。

3. 系统采用 Streamlit 实现快速开发与部署，为用户提供实时预测和直观数据展示。

a) 待改进点：

4. 数据来源相对单一，未能全面覆盖其余认定，限制了模型的普适性。

5. 系统部署的功能尚处于基础阶段，未涵盖用户管理、权限控制等企业级需求，后续可进一步扩展。

6.2 展望

在未来研究方向主要有以下两个方面：

1. 数据扩展与多维度分析：未来可通过收集更多地区、更多企业的历史数据以及多维度特征，进一步丰富预测模型的输入特征，提高模型的泛化能力和适用范围。例如，可以补充企业的财务指标、行业趋势数据以及政策环境变化等外部因素。同时，尝试对跨行业数据进行关联分析，以探索行业间的共同点和差异，进一步优化模型的预测能力。

2. 系统功能扩展：未来计划基于 Django 框架开发一个集成认定咨询、AI 问答、数据展示及评分预测等功能的多模块化综合平台。该平台将包含用户友好的界面，提供丰富的数据可视化功能和精准的预测工具，帮助用户直观了解企业的核心指标和评分结果。AI 问答模块将基于深度学习模型，支持多轮互动，解答用户关于认定流程、行业政策及数据分析的相关问题。同时，系统将支持数据上传、实时预测以及权限管理等企业级应用功能，为政策制定者、企业管理者和研究人员提供全方位的数据支持和决策辅助。

在将来，随着国家对“专精特新”中小企业支持政策的不断深化，企业评分预测系统在政策评估、投资分析和企业管理等领域将发挥重要作用。该系统不仅能够为中小企业的政策申报和认定提供数据支持，还可以帮助企业在竞争中识别优势和短板，优化发展战略。同时，政府和投资机构可以利用系统进行行业趋势分析、信用风险评估及资源配置优化，推动中小企业健康发展。

随着大数据技术的发展，该系统可以进一步整合在线数据源，实现更大规模的数据处理和实时分析。此将预测系统部署到移动端设备，可大幅提升用户的使用便捷性和覆盖范围。通过将人工智能与产业实际需求深度结合，该系统有望推动中小企业向智能化、数字化方向转型，并为国家经济高质量发展和结构升级提供有力支撑。

参考文献

[1]佚名.财政部等两部门:发文支持"专精特新"中小企业高质量发展[J].中国食品, 2021(5):2.

[2]李培恩.中小企业须走"专精特新"发展之路[J].化工管理, 2011(5):2.DOI:CNKI:SUN:FGGL.0.2011-05-010.

[3]刘昌年,梅强."专精特新"与小微企业成长路径选择研究[J].科技管理研究, 2015, 35(5):5.DOI:10.3969/j.issn.1000-7695.2015.05.024.

[4]刘晨.从"专精特新"到"隐形冠军"[J].企业家信息, 2024(4):5-7.

[5]何嘉."专精特新"中小企业高质量发展研究[J]. 2024.

[6]朱宏任.中国企业发展与企业家精神:从专精特新到世界一流[J].军工文化, 2024(1):67-68.

[7]本刊编辑部.航天品质 家国情怀——长城润滑油创建世界一流专精特新示范企业侧记[J].中国航天, 2023(7):69-70.

[8]乐烨.物流公司报关业务的专精特新——以上海金陵国际物流有限公司为例[J]. 2021.DOI:10.3969/j.issn.1674-4993.2021.01.015.

[9]武昭媛.专精特新"小巨人"企业股权融资情况分析[J].[2024-11-20].

[10]刘新荷.供应链金融视角下数字化对"专精特新" 中小企业绿色技术创新的影响研究[J].电子商务评论, 2024, 13(1):649-657.DOI:10.12677/ECL.2024.131075.

[11]胡海波,周洁.专精特新企业迭代的底层逻辑[J].企业管理, 2024(7):117-120.

[12] Cui T , Kerlin J A .China: The Diffusion of Social Enterprise Innovation: Exported and Imported International Influence[J]. 2017.DOI:10.1108/978-1-78714-250-320171004.

[13] Yi Z , Zhanbin F , Xi Z ,et al.Geographical Location,Environmental Regulation and Enterprise Innovation Transformation[J].Journal of Finance & Economics, 2016.DOI:10.16538/J.CNKI.JFE.2016.09.008.

[14] Foundation D S .Django: The Web framework for perfectionists with deadlines[J]. 2013.

[15] Can T .Review of next-gen of front-end Web interface technology[J].Journal of Chongqing Technology and Business University(Natural Science Edition), 2009.

致谢

致所有爱我的人！

Python 数据分析课程设计期末考核成绩及评语

姓名： 包旻宇 学号 221470101

班级 21 级数据科学与大数据技术（专升本）1 班

期末考核总分：

评语：

评阅教师：陶建敏

评阅时间：