编号： 202121702044

河北工业大学硕士学位论文

**基于集成学习的在线用户创新社区创意选择研究**

|  |  |  |  |
| --- | --- | --- | --- |
| 论 文 作 者： | 高如意 | 学 生 类 别： | 全日制 |
| 学 科 门 类： | 管理学 | 学 科 专 业： | 管理科学与工程 |
| 指 导 教 师： | 李向东 | 职 称： | 教 授 |

2023年12月

Thesis Submitted to

Hebei University of Technology

for

The Master Degree in

Management Science and Engineering

**Research on Creative Selection of Online User Innovation Community Based on Ensemble Learning**

By

Gao Ruyi

Supervisor: Prof. Li Xiangdong

Dec, 2023

**摘 要**

数字经济时代的到来标志着企业竞争格局的深刻变革，其中数字创新成为取得竞争优势的关键要素。在这一大背景下，在线用户创新社区（Online User Innovation Community，OIC）崭露头角，成为企业吸引外部创新资源、推动数字经济发展的重要手段。然而，不同OIC的运营效能存在明显的差异，有些成功运营，而有些却因失败或关闭而引起关注，即尽管创意收集是OIC的核心活动之一，但实际应用却常常缺失，导致创意的商业价值不够显著。管理层在面临信息筛选困难的情况下，往往难以找出高价值的创意，从而限制了企业在数字经济时代的竞争力。在应对OIC的管理挑战时，企业需要更注重实际应用和高效管理。因此，本文通过在线用户创新社区创意价值的研究，能够有效地了解创意以及创客信息对其影响的机制，为了更有效地选择高质量的创意构建了创意选择模型，提高对创意筛选的准确性和时效性，有助于企业更有针对性地投入资源，提高创新效率，从而在市场竞争中取得更大的优势。

本文首先通过梳理相关文献，总结开放式创新理论、在线用户创新社区概念、ELM模型、文本挖掘、Logistic回归模型、集成学习算法和贝叶斯优化相关理论和研究方法，为在线用户创新社区影响因素分析及创意选择模型构建提供理论和方法支持。根据ELM模型本文对在线用户创新社区创意价值的影响因素提出假设，然后选择Salesforce在线用户创新社区作为研究对象， 通过Python编写爬虫脚本对创意相关信息进行采集，并对数据进行预处理、描述性统计和相关性分析，再利用Logistic回归实证分析假设，回归结果证明所提出的标题详细度、标题可读性、内容可读性、标题与内容文本相似度、标题情感倾向、内容情感倾向、创意点赞数、创意得分数、创客徽章数、创客技能标签数都显著影响创意价值，不仅如此，创客的先前经验正向调节着创客徽章数对创意价值的影响。然后通过这些特征因素，利用RF、ET、GBDT、XGBoost、AdaBoost和LightGBM五类集成学习算法进行单独对社区创意进行筛选，研究表明XGBoost算法在五种算法在选择效果上较为突出，使之作为改进算法进一步对创意进行筛选，通过SMOTE方法对样本均衡处理以及贝叶斯优化XGBoost算法的参数，发现改进XGBoost算法在AUC值上比XGBoost算法提高了1.39%。最后根据回归和创意选择结果进行分析得出研究结论和启示，并总结了研究的不足以及对未来研究的展望。

**关键词：**在线用户创新社区；创意价值；ELM模型；创意选择；集成学习

ABSTRACT

The arrival of the digital economy era marks a profound transformation in the competitive landscape of enterprises, among which digital innovation has become a key element in obtaining competitive advantages. In this context, Online User Innovation Community (OIC) has emerged as an important means for enterprises to attract external innovation resources and promote the development of the digital economy. However, there are significant differences in the operational efficiency of different OICs, with some successfully operating while others attract attention due to failure or closure. Although creative collection is one of the core activities of OICs, practical applications are often lacking, resulting in less significant commercial value of creativity. In the face of difficulties in information screening, management often finds it difficult to identify high-value ideas, thereby limiting the competitiveness of enterprises in the digital economy era. When dealing with the management challenges of OIC, enterprises need to pay more attention to practical application and efficient management. Therefore, through the study of the creative value of online user innovation communities, this article can effectively understand the mechanism of the impact of creativity and maker information on them. In order to more effectively select high-quality creativity, a creative selection model has been constructed to improve the accuracy and timeliness of creative screening. This helps enterprises to invest resources more targeted, improve innovation efficiency, and gain greater advantages in market competition.

This article first summarizes the theories and research methods of open innovation theory, online user innovation community concept, ELM model, text mining, logistic regression model, ensemble learning algorithm, and Bayesian optimization by reviewing relevant literature, providing theoretical and methodological support for the analysis of influencing factors of online user innovation community and the construction of creative selection models. Based on the ELM model, this article proposes hypotheses on the influencing factors of creative value in online user innovation communities. Then, Salesforce online user innovation communities are selected as the research object, and Python is used to write web scraping scripts to collect creative related information. The data is preprocessed, descriptive statistics, and correlation analysis are performed. Logistic regression is used to empirically analyze the hypotheses, and the regression results prove that the proposed title detail, title readability, content readability, title to content text similarity, title emotional tendency, content emotional tendency, creative likes, creative score, maker badge number, and maker skill tag number all significantly affect creative value. In addition, the previous experience of the maker positively regulates the number of maker badges. The impact on creative value. Then, based on these characteristic factors, five integrated learning algorithms, RF, ET, GBDT, XGBoost, AdaBoost, and LightGBM, were used to individually screen community creativity. The study showed that XGBoost algorithm had outstanding selection performance in the five algorithms, making it an improved algorithm for further screening creativity. Through SMOTE method for sample balancing and Bayesian optimization of XGBoost algorithm parameters, it was found that the improved XGBoost algorithm increased AUC value by 1.39% compared to XGBoost algorithm. Finally, based on the regression and creative selection results, the research conclusions and inspirations were analyzed, and the shortcomings of the research were summarized as well as the prospects for future research.

**Keywords:** online user innovation community; Creative value; ELM model; Creative selection; Integrated learning

目 录

第一章 绪论 - 1 -

1.1 研究背景 - 1 -

1.2 研究目的与研究意义 - 2 -

1.2.1 研究目的 - 2 -

1.2.2 研究意义 - 3 -

1.3 研究内容 - 4 -

1.4 研究方法与技术路线 - 4 -

1.4.1 研究方法 - 4 -

1.4.2 技术路线 - 6 -

1.5 创新点 - 7 -

第二章 理论基础与文献综述 - 9 -

2.1 理论基础 - 9 -

2.2.1 开放式创新理论 - 9 -

2.2.2 在线用户创新社区概念 - 10 -

2.2.3 ELM模型 - 11 -

2.2.4 文本挖掘 - 12 -

2.2.5 Logistic回归模型 - 13 -

2.2.6 集成学习算法 - 13 -

2.2.7 贝叶斯优化 ..- 19 -

2.2 文献综述 - 20 -

2.2.1 在线用户创新社区相关研究 - 20 -

2.2.2 创意价值影响因素相关研究 - 22 -

2.2.3 创意选择方法相关研究 - 23 -

2.2.4 研究述评 - 24 -

2.3 本章小结 - 25 -

第三章 研究假设与模型构建 - 26 -

3.1 中心路径假设 - 26 -

3.1.1 创意文本类型特征 - 26 -

3.1.2 创意规范类型特征 - 28 -

3.2 边缘路径假设 - 29 -

3.2.1 创客特征 - 29 -

3.2.2 用户评论特征 - 30 -

3.3 模型构建 - 31 -

3.4 本章小结 - 32 -

第四章 研究设计与实证分析 - 33 -

4.1 研究对象与数据采集 - 33 -

4.2 变量获取 - 35 -

4.2.1 因变量 - 35 -

4.2.2 自变量 - 35 -

4.2.3 调节变量 - 39 -

4.2.4 控制变量 - 39 -

4.3 描述性统计与相关性分析 - 40 -

4.3.1 描述性统计 - 40 -

4.3.2 相关性分析 - 41 -

4.4 Logistic回归模型建立 - 43 -

4.5 Logistic回归结果分析 - 44 -

4.6 本章小结 - 46 -

第五章 在线用户创新社区创意选择方法研究 - 48 -

5.1 创意选择模型框架 - 48 -

5.2 集成分类器的评价指标与选择 - 49 -

5.2.1 评价指标 - 49 -

5.2.2 集成分类器的选择 - 50 -

5.3 改进XGBoost算法 - 53 -

5.3.1 基于SMOTE优化数据集 - 54 -

5.3.2 基于贝叶斯方法优化参数 - 55 -

5.4 在线用户创新社区创意选择结果 - 59 -

5.5 特征重要性分析 - 60 -

5.6 本章小结 - 62 -

第六章 结论与展望 - 63 -

6.1 研究结论 - 63 -

6.2 管理启示 - 65 -

6.3 研究局限性及研究展望 - 67 -

参考文献 - 68 -

图和附表清单

**图清单**

[图 1.1技术路线图 - 7 -](#_Toc154583588)

[图 2.1. ELM模型 - 11 -](#_Toc154583589)

[图 3.1 在线用户创新社区创意价值影响因素研究模型 - 31 -](#_Toc154583590)

[图 4.1 Salesforce在线用户创新社区界面（部分） - 33 -](#_Toc154583591)

[图 4.2 Salesforce 创意特征示例 - 34 -](#_Toc154583592)

[图 4.3 Salesforce 创客特征示例 - 34 -](#_Toc154583593)

[图 4.4 变量相关性热力图 - 43 -](#_Toc154583594)

[图 5.1 在线用户创新社区创意选择模型框架 - 49 -](#_Toc154583595)

[图 5.2 集成学习模型的ROC曲线图 - 52 -](#_Toc154583596)

[图 5.3 改进XGBoost算法过程图 - 54 -](#_Toc154583597)

[图 5.4 XGBoost参数优化效果 - 58 -](#_Toc154583598)

[图 5.5 优化的XGboost模型与XGBoost模型的ROC曲线对比 - 60 -](#_Toc154583599)

[图 5.6 在线用户创新社区创意选择的特征重要性 - 61 -](#_Toc154583600)

**表清单**

[表 4.1 模型变量与定义 ……………………………………………………………….- 39 -](#_Toc154583687)

[表 4.2 描述性统计分析 - 40 -](#_Toc154583688)

[表4.3 变量之间相关性 - 41 -](#_Toc154583689)

[表 4.4 Logistic回归结果 - 44 -](#_Toc154583690)

[表 4.5 研究假设结果汇总 - 46 -](#_Toc154583691)

[表 5.1 混淆矩阵 - 49 -](#_Toc154583692)

[表5.2 集成分类器在线用户创新社区创意选择效果 - 51 -](#_Toc154583693)

[表 5.3 XGBoost集成学习算法参数空间定义 - 56 -](#_Toc154583694)

[表 5.4 贝叶斯优化参数初始值 - 56 -](#_Toc154583695)

[表 5.5 贝叶斯改进XGBoost算法伪代码 - 57 -](#_Toc154583696)

[表 5.6 XGBoost对在线用户创新社区创意选择的最优参数 - 58 -](#_Toc154583697)

[表 5.7 改进XGboost模型与XGBoost模型的创意选择效果对比 - 59 -](#_Toc154583698)

[表 5.8 改进XGBoost集成学习算法后的特征重要性结果 - 60 -](#_Toc154583699)

# 第一章 绪论

## 1.1 研究背景

作为提升企业核心竞争力的重要战略途径，创新越来越受到企业和研究者的重视。一直以来，人们认为创新是在生产或者服务企业内部独立完成的，但随着商品经济的发展，这种闭门造车式的创新难以应对日益变化的竞争环境及用户需求，企业逐渐认识到外部资源（如科研机构、供应商、用户等）对于创新的重要性，企业创新逐渐由封闭式向开放式发展[1, 2]。近年来，用户创新已成为一种重要趋势，用户在诸多领域的创新活动中都扮演着日益重要的角色[3, 4]。许多企业开始将用户创新纳入到企业的创新体系中，以提升生产效率、降低产品开发周期和生产成本[5]。 然而，传统的创新环境限制了用户的创新手段以及创新途径。一方面，时空上的隔阂使得用户无法参与企业的产品研发[6]；另一方面，由于缺乏行之有效的管理，用户的创新潜能很难在企业的创新活动中得到充分发挥[7]。不仅如此，在传统创新模式下，用户与用户之间难以互动与合作，这极大影响了用户的创新积极性。信息技术的飞速发展为用户创新提供了更为开放的平台，众多企业纷纷建立了自己的在线用户创新社区（Online user Innovation Community，OIC），以将用户创新纳入到自己的创新体系中来[8]。通过在线用户创新社区，用户可以更加方便地参与有益于产品创新的活动，例如，用户可以通过多种多样的数字化设备（如智能手机、平板电脑等）参与企业创新，创造和分享与创新相关的知识和经验[9, 10]。因此，在线创新社区得到众多学者的关注，被视为产品和服务创新的一个重要途径[11]。许多公司，包括IBM、奥迪汽车公司、微软、宝马和诺基亚等，都在不断探索如何能更高效地从在线创新社区中吸取并利用这些宝贵的知识和经验[12]。诸多企业案例表明，在激烈的市场竞争中，让用户参与创新能够使企业更加高效、准确地识别出客户的需求，提高企业的盈利能力[13]。

虽然在线用户创新社区在产生有价值的创意方面具有巨大的潜力，但是当用户贡献过载时，尤其是对于以文本格式提交的创意来说[14]，由于企业可用的资源和能力有限，很难评估如此大量且复杂的创意，且筛选出好的创意融入到自身产品开发中[15]。众包和在线用户创新社区都是一种开放式创新模式，在2006年，超过14万名国际参与者参加了IBM的创新活动，在一次众包竞赛中提交了46000多个创意，由于这些创意是由不同背景和专业程度的人提交的，所以众包中的文本数据通常包含大量的噪音和歧义。因此人工评估数据和过滤低质量创意的过程艰巨而漫长，甚至浪费大量成本和时间。但是企业及时的创意筛选行为有利于提升用户持续参与的主动性，提高用户未来创意的价值[16]，因此有必要通过研究为企业实现高效率地创意筛选提供指导意见。研究表明创意价值是创意未来市场接受度的可信指标，能够反应用户的需求信息[17]，从而降低企业创意选择的不确定性。

综合以上研究背景，尽管企业为了收集用户创意和获取外部创新构建了在线用户创新社区，但社区由于在创意选择过程中费时费力未能为企业带来理想的收益。因此，需要以整体化系统化的方法来看待在线用户创新社区的创意选择过程。为了在信息超载的环境下高效率地筛选出高质量的创意，首先分析影响创意价值的影响因素，然后基于这些影响因素构建自动化创意选择系统，目的更好地降低创意评估的工作量，更有效地利用外部的创意资源，从而为企业内部创意管理、开发提供可靠的依据。

## 1.2 研究目的与研究意义

### 1.2.1 研究目的

本文在自然语言处理技术基础上，通过对创意文本类型数据进行量化分析，深入探讨在线用户创新社区中创意价值的影响因素。并将其转化为集成学习的特征指标。进一步，将构建在线用户创新社区创意选择模型，以此为基础实现对新创意的自动化筛选，从而构建一个智能、高效的创意选择系统。该系统将不仅提升创意社区的运作效率，更有助于发现和推广具有实质价值的创意，本研究期望为创新社区管理和数字创意领域的发展贡献理论和实践的新思路，推动相关领域的进步与创新。

### 1.2.2 研究意义

（一）理论意义

研究视角进一步细化，从创意视角与创客视角结合。探索两个视角的特征对创意价值的影响。随着自然语言处理技术的发展和文本挖掘研究的进步，对文本的深入研究已成为在线用户创新社区的研究趋势。尽管在社交媒体、在线知识社区、在线评论社区等多种在线用户在线社区的环境下的研究使用文本挖掘获取了大量信息，但是在在线用户创新社区中尚未得到广泛应用，对非结构化创意文本数据的应用仍有不足。本文使用在线用户创新社区非结构化文本数据，通过自然语言处理的文本挖掘技术，从创意文本角度进一步挖掘创意价值的影响。本研究不仅丰富和扩展了国内外关于用户创新领域的相关研究，也为企业在实践过程中管理和挖掘用户创新模式的潜在价值提供了有益的帮助。

本文提出了一种基于集成学习原理的方法，以自动化创意选择过程。通过利用真实的互联网数据建立创意筛选系统，运用各种集成学习算法在在线用户创新社区上进行实证研究，从而有助于扩大该方法的应用领域，本文旨在为企业提供更有效的创意管理工具，使其能够更快速、智能地识别和采纳高质量的创意，从而提高创新效率和产品质量。

（二）实践意义

从创客角度而言，首先创客可以更加明确什么因素会导致自己的创意对企业更有帮助，企业也可以高效地选择出有价值的创意，减少自身选择的盲目性；企业可以及时了解创客需求，提高创客产品或者服务的使用体验，同时社区上企业和其他创客的及时反馈，也有助于提高创客满意度和忠诚度，促进创客持续贡献。

从企业角度而言，通过研究创意价值的内在机理，从用户创意文本中可以获取用户对产品和服务的需求，有助于企业完善平台管理机制以及引导用户产生高质量创意，为企业调整后续战略方针提供指导意见；通过构建创意选择模型，帮助企业在低人力成本下自动高效地筛选出高质量创意，对创意作出快速地反馈，促进产品的改进和研发，使在线用户创新社区更好地为企业创新服务，加快创新步伐，节省创新时间和成本，为企业后续可行性分析提供依据。

## 1.3 研究内容

（一）研究假设与模型构建

本章基于开放式创新理论，在 ELM模型基础上，与创意直接相关的特征作为中心路径假设，包括标题详细度、内容详细度、标题可读性、内容可读性、文本相似度、标题情感倾向、内容情感倾向、创意得分数以及创意点赞数；与创意间接相关的特征作为边缘路径假设，包括创客徽章数、创客技能标签数、创客先前经验以及用户评论情感倾向。分析其对创意价值的影响因素，提出对应的假设并且构建理论模型。

（二）研究设计与实证分析

本部分以Salesforce在线用户创新社区上的创意为研究主体，通过数据收集与预处理，采用自然语言处理技术对相关变量进行测量，然后进行数据描述性统计与相关性分析、模型检验与结果分析等过程进行实证分析，研究在线用户创新社区创意价值的影响因素，得出相应结果和研究结论。

（三）在线用户创新社区创意选择方法研究

本章在上一章得到显著性影响因素的结果基础上，获得集成学习的特征指标。基于集成学习构建在线用户创新社区创意选择模型，通过不同集成学习算法性能的对比，综合准确度和运算效率选择较优的模型，并通过优化方法进行模型改进，实现在线用户创新社区创意的选择，并探究选择模型中各特征的重要性排序，为社区建立自动化创意筛选系统提供解决方案。

## 1.4 研究方法与技术路线

### 1.4.1 研究方法

（1）文献分析法

本文采用了多渠道获取信息的策略。通过充分利用学校图书馆、百度学术、中国知网、SCI-HUB、Web of Science等专业资源，广泛涵盖国内外最新文献。在信息的汇总过程中，运用了系统性的归纳与演绎方法，以明确研究的理论支撑、前沿动态，以及未来的发展趋势。在梳理文献的同时，特别注重对模型变量含义的精确定义，确保研究框架的清晰度和稳固性。这一全面而深入的文献综述为研究提供了坚实的基础，使得在理论构建和研究设计上更具有权威性和可靠性。通过对各种资源的综合应用，本文力求为在线用户创新社区和创意价值主题的深入探讨提供了全面而可靠的知识支持。

（2）文本挖掘法

文本挖掘是针对半结构化或非结构化自然文本的数据挖掘过程，旨在发现数据中的潜在特征。本文在分析创意文本方面，文本挖掘技术能够系统地识别每个创意的特征，为企业提供深入的了解。通过对文本挖掘的结果进行分析，企业能够获得关键信息，从而优化产品条件和平台管理，提高竞争力。文本采用文本挖掘技术深入挖掘创意文本，提取相关特征，并将这些信息量化为自变量，以进行统计分析和预测。这一方法有助于企业更全面地了解创意的本质，而且为决策者提供了基于数据的依据，以更有效地制定战略和决策，从而推动企业在市场中取得成功。

（3）实证研究法

基于现有理论，本研究选取Salesforce在线用户创新社区作为研究对象，采用P编写爬虫程序获取Analytics板块的CRM Analytics、operational Dashboards与Operational Reporting三个创意类型各项数据所发布的所有的创意项目数据，共获取12558条创意相关内容，然后对数据进行预处理，大致了解数据的分布状况。提取创意和创客的结构化特征以及从创意文本中挖掘文本特征，通过实证分析方法对在线用户创新社区中创意价值的影响因素进行研究，探究其对创意价值的影响机理。在此基础上利用集成学习算法进行创意选择分析，建立创意选择模型，为企业提高创意筛选效率提供依据。

（4）统计分析法

本文在实证研究过程中利用SPSS软件对提取的数据特征进行分析，对相关研究变量进行描述性统计分析和相关性分析，获取各研究变量的数值分布情况以及研究彼此之间的相关程度和相关方向，然后进行逻辑回归分析，构建在线用户创新平台创意价值影响因素模型，探究文本特征和结构化特征对创意价值的影响因素。

（５）集成学习法

本研究基于提取的特征，利用多种集成学习算法，构建在线用户创新社区创意模型，实现创意筛选的自动化，提高企业创新能力。

### 1.4.2 技术路线

根据本文的研究内容和研究方法画出如下技术路线，如图1.1所示：

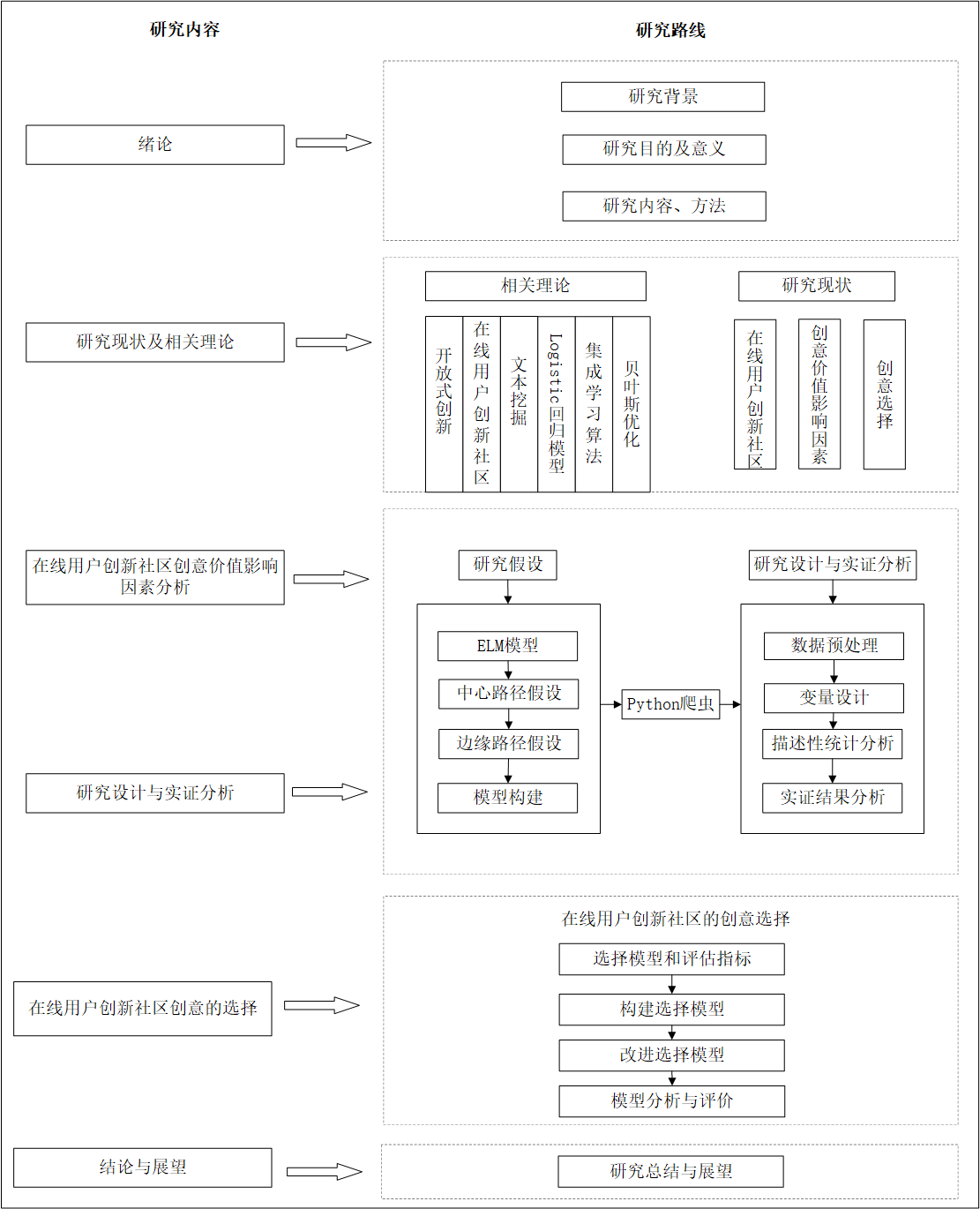


图 1.1技术路线图

Figure 1.1 Technical roadmap

## 1.5 创新点

（1）构建了在线用户创新社区创意价值的影响因素实证模型。

在信息过载的条件下，在线用户创新社区的潜能是多种多样的，但系统地识别社区中创意价值的问题很少被关注。本文基于ELM模型，分别创意文本类型特征、创意规范类型特征、创客特征以及用户评论特征四个维度提出影响创意价值的研究假 设。基于Salesforce在线用户创新社区中的创意进行实证分析。全面探讨不同特征因素对创意价值影响的作用机理，弥补现有相关文献中的不足。研究结果对在线用户创新社区管理者对创意价值识别和搜索提供管理启示，对社区用户提高创意被企业实施的可能性提供参考策略。

（2）构建在线用户创新社区创意选择模型，提高创意筛选效率。

原有研究多是基于统计推断模型探究影响因素有哪些，虽然研究结论比较合理，但未从本质上解决人工筛选创意时效率低下的问题，如果创意反馈时间过长，则会降低用户发布创意的积极性和活跃性。因此本文结合结构化特征和创意文本，基于机集成学习算法，建立在线用户创新社区创意选择模型，实现自动化且高效率地筛选高质量创意，促进了信息传播。本文利用网络上的大数据进行研究分析，避免人工评定的主观性，数据更加客观、可信，扩展了在线用户创新社区中创意筛选的相关研究。

（3）改进XGBoost算法，提高了在线用户创新社区创意选择模型的综合性能。

本文首先利用SMOTE算法对不具有创意价值的样本数量与具有创意价值的样本数量严重不均衡的问题进行了采样优化，然后利用贝叶斯优化算法和10折交叉验证得到了相比传统XGBoost算法对创意选择的效果更好的改进XGBoost算法，利用这一优化后的模型得到在线用户创新社区的相关指标的特征重要性及排序，更好地解释创意价值的关键影响因素，提升了模型对创意选择的应用效果。

# 第二章 理论基础与文献综述

## 2.1 理论基础

### 2.2.1 开放式创新理论

随着全球化的发展以及市场竞争力日益激烈，在知识快速扩散的时代，企业传统的封闭式创新模式已经不能满足市场竞争的需求，因此新的创新形式由此诞生。开放式创新理论最早由Chesbrough首次提出[18]，强调企业突破组织边界，利用和整合外部资源转化为企业自身的创新能力，在此基础上，陈劲等[19]认为随着经济全球化的不断深入，企业边界变得模糊，企业利用和整合外部资源的能力成为企业创造价值的重要来源。根据创新知识的流动方向，可以将开放式创新分为三种类型：内向型开放、外向型开放和混合型开放。其中内向型开放是指创新资源从外部流入企业内部从而提高企业内部创新能力，外向型开放是指企业将自身的资源和技术输出到外部环境从而获取商业利益，混合型开放是指开放模式没有固定的属性，而是根据企业的发展目的进行多种模式的交叉融合，既有知识的流入又有资源的流出[20]。依据调查研究发现，内向型开放式创新模式是企业目前实施较为普遍和规范化的一种模型，而外向型开放式创新模式只在部分企业实施，还处于初步发展的阶段[21]。

Muller等人认为开放式创新过程分为三个阶段，分别为：创意产生阶段、创意发展阶段和创意商业化阶段[22]。创意产生阶段主要是指企业尽可能多的收集用户创意，获取市场需求信息，然后识别出价值高的关于产品和服务的创意。创意发展阶段主要任务是将创意转化为市场机会，创意商业化阶段是指基于新的创意调整企业发展模式以获得更多的市场机会。尽管开放式创新不断发展，但是在开放式创新过程中仍然存在一定的挑战和风险亟待解决，例如高质量创意识别，创意实施中的资源、市场、分销等各种问题。

### 2.2.2 在线用户创新社区概念

在线用户创新社区是企业为了吸引用户参与企业内部创新的虚拟创新社区，在线用户创新社区吸引了大量的善于创新的用户，他们可以在在线用户创意社区自由提交关于产品、流程等各种创意[23]，企业利用在线用户创新社区用户的创意和想法，可以择取有价值并且有实施意义的创意，将其利用在自身产品的改进和创新中，进而将企业的创新能力提高[24]，秦敏认为在线用户创新社区是在线用户直接或者间接参加企业内部的研发等与创新有关的在线活动的社区，并且将在线用户创新社区分为用户创新社区和开放源代码创新社区[25]。在本文中，在线用户创新平台主要是指用户创新社区而非开放源代码社区，是外部用户通过企业构建的平台为企业内部产品创新提供建议的平台。

随着在线用户创新社区的发展以及众多学者对在线用户创新社区的研究，在线用户创新社区已经被越来越多国内外企业所应用：例如dell、salesforce、lego等国外企业建立了较完善的在线用户创新平台，而国内在线用户创新平台发展较晚，只有海尔、小米等少数企业构建了在线用户创新社区，且仍处于初步发展阶段。社区、用户、创新成果和制度是企业在构建或运营在线用户创新社区时需要考虑的关键因素，创建创新社区和社区外用户的参与，促进企业创新的持续互动，有效的社区信息管理和创新成果管理，完善的社区系统将促进用户的积极参与，为社区创造良好的创新环境[26]。

因此，关于在线用户创新社区的研究可分为三大角度：一是从社区平台自身出发，研究企业在构建在线用户创新社区时需要具备的可实施性条件并对企业起到什么作用；而是从社区平台的在线用户出发，研究其在社区平台中的参与创新因素以及用户分类等；三是从在线用户在社区平台发布的创意角度，研究其创意采纳影响因素以及创意识别管理等。

### 2.2.3 ELM模型

精细加工可能性模型（The Elaboration Likelihood Model，ELM）由Richard Petty 和John Cacioppo在1981年提出的一种处理说服过程的心理学理论[27]，描述信息的特征如何影响一个人态度的形成，进而影响其行为转变[28]。

ELM型基于信息加工深度和个人主观精确性需求的不同，解释说服有两种路径，即中心路径和边缘路径[29]。中心路径是指人们对信息进行精密的思考从而改变态度的过程，主要涉及与信息质量相关的信息内容的特征[30]，如信息内容的客观性和丰富度等；边缘路径是指人们没有足够的能力处理信息内容，可以通过简单的线索和边缘信息，不涉及深入的思考而改变态度的过程[31]，可以与信息源特征相关，也可以与信息相关，如信息源的可信度和权威性等都可以作为外围线索引起态度和行为的改变，信息的长度或者逻辑论证中观点的数量也可以作为外围线索影响态度转变。通过中枢路径处理信息比通过边缘路径处理信息需要更多的认知成本，因此在中枢路径上态度的形成和转变更加稳定持久，而在边缘路径上的态度和行为的变化相对短暂[32]，但是当人们缺乏时间和精力处理信息时，边缘路径上的影响因素对人们态度的转变有着重要的作用[33]，模型如图2.1所示。

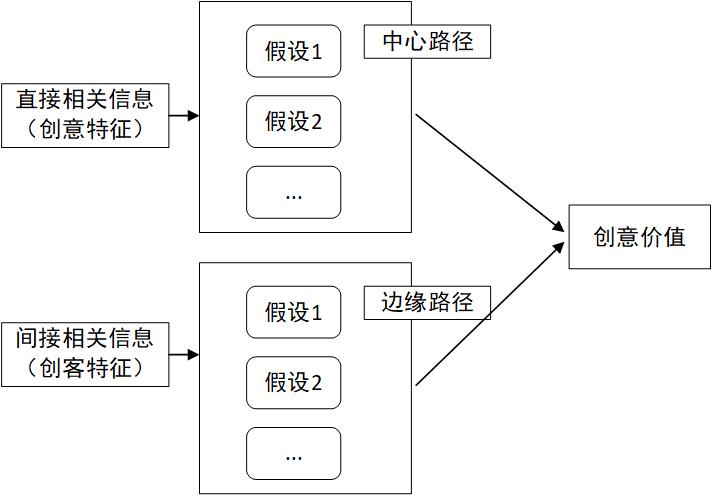


图 2.1. ELM模型

Figure 2.1 ELM model

### 2.2.4 文本挖掘

在信息共享快速发展的时代，越来越多的用户个体在网络上分享各种观点和想法，构建了一个庞大而丰富的信息共享平台[34]。这些网络内容涵盖了广泛的主题，反映了用户的关切和兴趣，同时也为研究者提供了丰富的研究资源和机会[35]。然而，随着信息爆炸式增长，这些网络内容的大部分超过80%是非结构化的，包括文本、视频、图像和音频等多种形式。这使得传统的研究方法面临着前所未有的挑战，尤其是在对这些庞杂的、非结构化的数据进行分析和研究。在过去，研究人员通常采用定性数据分析方法，例如手动编码对自然语言进行处理，以理解和挖掘用户观点和想法。然而互联网产生的数据量呈指数级增长，超越了传统的人工分析的处理能力。在这种背景下，为了更有效地处理大规模、非结构化的数据，文本挖掘方法应运而生。

文本挖掘技术是一种通过统计和机器学习等手段，从非结构化文本中提取有用信息的方法[36]。其核心目标不仅在于进行文本的计算和定量分析，更在于揭示文本背后隐藏的数据，以解决传统研究设计中对边缘化群体的忽视[37]。该技术主要应用于处理半结构化或非结构化数据，运用文本结构分析、文本摘要、文本分类、文本聚类、关联分析、情感和趋势预测等多种技术方法[38]。在实践中，文本挖掘涉及众多领域，包括但不限于信息技术、文本分析、统计学、数据可视化、数据库技术和机器学习。整个文本挖掘过程包括几个关键步骤：首先，需要获取文本数据集；其次，对数据进行预处理，以清理、标准化和转换文本数据，为后续分析做好准备；然后，通过各种技术手段进行文本挖掘分析，其中涉及到主题挖掘、情感分析、信息检索、机器翻译、自动文摘、信息过滤和信息安全等多个领域的应用；最终，对挖掘结果进行评估，确保其质量和可靠性。

文本挖掘技术的广泛应用使其成为一个跨学科的研究领域，不仅拓展了信息科学的边界，也为解决社会问题和改善决策提供了强有力的工具。未来，随着技术的不断发展和创新，文本挖掘有望在更多领域展现其潜力，并为社会发展和科学研究带来新的可能性。

### 2.2.5 Logistic回归模型

在Logistic回归模型中，因变量（被解释变量）只能取两个离散值，通常表示为1和0。Logistic回归分析方法根据被解释变量的取值类别可以分为二项Logistic回归（二值Logistic回归、二分类Logistic回归）分析和多项Logistic（多值Logistic回归、多分类Logistic回归）分析。在二项Logistic回归模型中，因变量仅能取两个值，分别用1和0表示。

通常用P表示因变量为1的概率（事件发生的概率），1-P表示因变量为0的概率（事件不发生的概率），如公式（2.1）所示：

当F为线性函数时，P可表示为：

但是，由于P的值在[0, 1]区间内，当P接近于0或1时，尽管自变量发生了很大变化，P值的变化也不会太大，因此使用线性模型对（2.1）进行参数估计是不可行的。函数P对X和r的变化，在P接近于0或1时是不敏感且变化较缓慢，呈非线性。因此，需要寻找一个函数F(P)，使得该函数在P接近于0或1时有较大幅度的变化，但函数形式又不过于复杂[85]。所以引入Logit变换：

通过Logit变换后，再利用一般线性回归建立Y与X, 之间的线性关系模型：

在完成参数估计、模型评估以及模型检验后，由公式（2.4）可知，可以预测给定样本为1的概率P的预测值：

### 2.2.6 集成学习算法

机器学习领域自问世以来，相关研究人员一直在寻求不断改进和提高模型的性能，经历了许多重大的发展和变革，集成学习在机器学习领域中占据了极为重要的地位。集成学习通过结合多个不同的学习器，可以降低过拟合的风险。每个基本学习器可能对数据的不同方面有着不同的见解，通过结合它们的输出，可以减轻模型对训练数据中的噪声和降低过拟合的风险，从而提高泛化性能。这种多模型的组合方式可以作为一种模型多样性的引入，有助于提高整体性能。此外，不同算法可能在不同方面表现出色，而集成学习能克服单一算法的局限性。

**（1）随机森林**

随机森林（Random Forest，RF）常用于分类和回归任务。它是一种集成学习方法，通过组合多个决策树来提高模型的性能和鲁棒性。RF由Tin Kam Ho于1995年首次提出，后来由Leo Breiman在2001年进行了进一步的改进。

RF的核心组成部分是决策树。决策树是一种树形结构，用于根据输入特征对数据进行分类或预测，每个节点代表一个特征，每个叶节点代表一个类别或一个数值。在构建决策树时，算法会选择最佳的特征进行数据分割，目的使分类误差或均方误差最小化。RF还引入了特征随机性，它在每个节点上只考虑一个随机子集的特征，而不是考虑所有特征。这有助于减少特征间的相关性，使每棵树更加独立，从而提高了模型的多样性。

在分类任务中，随机森林使用投票机制来确定最终的类别。每棵决策树都对输入数据进行分类，然后选择获得最多投票的类别作为最终预测结果。在回归任务中，随机森林取所有决策树的平均值作为最终预测。随机森林通过引入随机性来增加模型的多样性，减少过拟合风险，提高泛化能力。由于它的随机性成分，随机森林通常不需要太多的超参数调优，相对容易使用。

**（2）梯度提升决策树**

梯度提升决策树（Gradient Boosting Decision Trees，GBDT），用于解决回归和分类问题，通过逐步构建决策树来提高模型性能。与传统的梯度提升方法不同，GBDT以决策树为基本模型，每一次迭代GBDT会训练一棵新的决策树，然后将其与之前的树进行组合。这意味着GBDT是一个增量建模过程，每次迭代都会学习一个新的决策树来纠正之前模型的残差，算法流程如下：

假设给定一个二分类的训练数据集T。

1. 初始模型是对数比率的初始值，通常是类别的先验概率的对数比率：

每轮迭代进行以下步骤：

1. 计算负梯度（残差），计算损失函数对模型输出的负梯度。

其中，是样本i在第m-1轮迭代后的概率预测。

1. 拟合决策树，使用残差拟合一个决策树，表示为。
2. 计算决策树的输出，计算决策树的输出，通常为叶节点的对数比率的平均值。

其中，是样本i被分配到叶节点的权重，λ是正则化项。

1. 更新模型的输出。
2. 计算样本概率，根据更新后的模型输出计算样本属于类别 1 的概率。

③最终模型，为所有基本决策树的加权和。

**（3）AdaBoost**

Adaboost是一个迭代算法，它会多次训练弱分类器，并根据每个弱分类器的性能来更新样本的权重，Adaboost的主要思想是不断调整样本的权重，使模型更加关注那些难以分类的样本，从而提高整体性能。它在实际应用中通常表现出色，特别是在处理二分类问题时。如果弱分类器过于复杂或数据中存在噪声，Adaboost可能容易过拟合。因此，合理选择弱分类器和进行交叉验证是使用Adaboost时的重要考虑因素。算法流程如下：

假设给定一个二分类的训练数据集。

①初始化训练数据（每个样本）的权值分布：如果有N个样本，则每一个

训练的样本点最开始时都被赋予相同的权重。

其中 i 表示第 i 个样本，N 表示样本数量。

②迭代训练

1. 训练弱分类器：在每轮迭代中，训练一个弱分类器
2. 计算分类误差率:

其中，yi表示第 i 个样本的真实标签，表示弱分类器在第 t 轮迭代中的输出，I(·)是指示函数，等于1当括号内的条件成立，否则等于0。

c. 计算弱分类器的权重：

d. 更新样本权重，对于每个样本：

e. 归一化样本权重，对所有样本i 进行归一化处理：

③构建强分类器

最终的强分类器 H(x) 由所有弱分类器的加权组合构成：

其中sign(·)的符号函数，如果括号内的值大于等于0，则输出+1，即正类，否则输出-1，即负类。

**（4）XGBoost**

XGBoost是基于决策树模型的梯度提升算法，假设在线用户创新社区的创意数据集为E={(xi，yi) | xi∈Ｒm，yi∈Ｒ}，其中Ｒm为特征空间，数据子集xi对应的标签为yi，则T棵回归树最终的创意价值预测结果为：

Q为第i个样本最终的预测结果，ft为独立的回归树函数，F为树的集合，ft(xi)为第t棵回归树上xi样本的预测值。对目标函数定义如下:

Obj为损失函数；yi为第i个样本集的真实标签；为损失函数；T为叶子节点的数量； 为第j个叶子节点的权重；γ和λ为惩罚项，控制模型的复杂度。

通过加入二阶泰勒展开式，进行k次迭代训练，可改写目标函数如式（）所示：

gi为损失函数的一阶导函数，Ij表示属于第j个叶子节点的所有样本xi划入到一个叶子节点中的样本集合，hi为损失函数的二阶导函数。

对式求偏导，将等于0时的wj值代入得到式(4)，以该式判断树的结构好坏。

式中的值越小代表树的拟合精度越好。

**（5）LightGBM**

LightGBM采用了基于直方图的学习算法，通过一系列创新性的设计来加速梯度提升树的训练过程。首先，LightGBM使用了基于直方图的数据存储和分割算法，将连续特征离散化成直方图形式，有效减少了内存占用和计算复杂度。这使得LightGBM能够高效地处理大规模数据集，同时能够处理高维稀疏特征。在每轮迭代中，LightGBM采用了Leaf-wise生长策略，与Level-wise生长相比，Leaf-wise生长更加高效。该策略在每次选择分裂节点时，优先选择能够最大程度降低损失函数的节点进行生长。这使得LightGBM能够更快地找到最优的分裂点，加速模型的收敛过程，算法流程如下：

假设给定一个二分类的训练数据集。

①初始模型是对数比率的初始值，通常是类别的先验概率的对数比率。

②对每轮迭代进行以下步骤：

1. 计算损失函数对模型输出的负梯度（一阶导数）和二阶梯度。

其中，是样本i在第m-1轮迭代后的概率预测。

1. 构建树，LightGBM使用负梯度和二阶梯度来构建一个决策树，以最小化当前节点的对数损失。
2. 计算决策树的输出，计算决策树的输出，通常为叶节点的对数比率的平均值。

其中，是样本i被分配到叶节点的权重，λ是正则化项。

1. 更新模型的输出。

其中， 是学习率，是指示函数，表示样本 x 是否在叶节点 j 中。

③最终模型，为所有基本决策树的加权和。

### 2.2.7 贝叶斯优化

贝叶斯优化是一种用于优化黑盒目标函数的迭代过程，其核心思想是利用贝叶斯统计推断来建模目标函数的潜在分布，并在每次迭代中选择最有希望的参数进行评估。这种方法尤其适用于那些目标函数计算代价昂贵或无法直接求导的情况。

①建模目标函数

假设有一个黑盒目标函数，其中x是输入参数（超参数），目标是在给定有限的预算内找到最小化或最大化这个目标函数的输入参数。

②选择先验分布

对于每个输入参数x，引入一个先验分布，假设为参数的分布空间。通常选择高斯分布作为先验分布，因为它的数学性质使得推断相对容易。

③更新先验分布为后验分布

使用贝叶斯定理，根据观测到的数据更新先验分布，得到后验分布。后验分布反映了对参数的最新结果，同时考虑了先验信息和观测数据，公式如下：

其中，D是观测到的目标函数值。

④建模目标函数的未知性： 在贝叶斯优化中，目标函数被建模为一个随机过程，通常是高斯过程。高斯过程是一个概率分布，对于任意输入参数 x给出一个关于目标函数值的概率分布。

⑤选择下一个观察点： 选择下一个输入参数以在目标函数上进行评估。

⑥观察目标函数： 在选择的输入参数处评估目标函数，得到相应的观测值。

⑦更新观测集合： 将新的观测值加入已有的观测集合，扩充数据集。

⑧迭代： 重复步骤 3 到 7 直到满足停止条件，例如达到最大迭代次数或观测到足够的数据。

通过不断更新先验分布、选择下一个观察点、观察目标函数，并迭代这一过程，贝叶斯优化逐渐收敛到目标函数的最优值，同时充分考虑了不确定性，使其在高维、噪声较大的优化问题中表现良好。

## 2.2 文献综述

### 2.2.1 在线用户创新社区相关研究

（一）创意方面相关研究

现如今阶段，企业封闭式创新已经无法满足现在市场竞争趋势，企业开始寻求在线用户创新方式，创建在线用户创新社区，所以企业要考虑自身情况是否有能力建立一个在线用户创新社区，并且要想创建一个完善体系和制度需要一些什么样的因素。有大量学者开始对企业合理构建在线用户创新社区以及如何吸收创新知识进行了研究，研究结果显示：在线用户创新社区的自身具有的网络外部性、开放性以及协同创新性，都对企业对自身产品的创新具有深远意义。

Hienerth等通过实证分析，发现在线用户创新社区的创意的创新效能比企业内部员工创新效能多3倍，企业在只需在创新社区投入少量成本就可以获得巨大的创新收益[39]；Ye和Kankanhalli以组织角度出发，实证分析了企业建立在线用户创新社区的目的：品牌竞争力和成本降低都显著正向影响企业建立在线用户创新社区，但前期准备成本以及创意价值评估成本显著负向影响企业建立在线用户创新社区的积极性[40]；Dong和Wu通过对星巴克企业实证分析，认为企业对社区创意积极采纳可以增加社区在线用户的活跃性，并且有利于提升企业价值[41]。李奕莹等基于创新价值链视角，通过仿真模拟发现用户个性化需求、用户奖励水平、用户生成内容以及企业动态能力是在线用户创新社区成功运营的关键影响因素[42]。吴冰等抓取小米在线用户创新社区相关数据进行实证研究，研究发现社区组织机制显著正向影响社区社会网络结构，且社区技术机制显著正向影响社区社会网络结构[43]。

虽然有很多企业成功地将在线用户创新理论运用到了实践中，但也有很多创新社区是不成功的。企业可能会因为对创新社区的管理能力不足，导致大量有可实施价值的创意未被发现等问题[44]。要想成功建立完善的社区并将创意有效地吸收需要面临四大挑战：（1）准确定位创意含义，及时审批；（2）设计有效的识别系统，准确决策有价值意义的创意；（3）社区平台要准确公布创意需求以及详细信息；（4）社区的健康可持续发展[45]。

（二）用户方面相关研究

在线用户创新社区中的在线用户是企业外部吸收产品创新信息的主要来源，而企业内部人员作为创新信息的吸收者，需要推动企业创新以及协调成员之间的关系[46]。小米公司在国内的在线用户创新社区建立中称得上一个典范：不管是产品功能的构想，还是新技术的研发，小米公司都充分地利用了创新社区在线用户的创意智慧[47]。因此，用户作为企业的最终消费者，其对企业的产品创新具有关键的作用，其也是企业产品的体验者，产品具有什么问题，都是最终消费者最先发现，所以其建议更能反应市场真实需求[48]。

Chen等（2012）通过对戴尔平台分析以及基于价值共创理论，研究了用户与用户之间交互性以及企业与用户之间的交互性对用户的创意量和质量的影响[49]。Bayus（2013）研究得出戴尔平台的在线用户的过去参与创新程度影响着其持续参与创新[16]。Lee和van Dolen（2015）研究发现创新社区的管理方式、工作人员的工作态度以及对问题反馈的速度都影响着在线用户参与创新创意的积极性[50]。Recker和von Briel（2017）通过分析某个企业创建在线用户创新社区失败的案例，发现创建在线用户创新社区缺乏大量在线用户的持续贡献是主要原因[51]。Prikkalainen等（2018）从心理学角度发现在线用户参与知识贡献动机受知识心理所有权影响[52]。李悦（2019）认为激励机制对用户创新动机有显著影响，通过构建创新意愿影响因素假设模型与实证分析，验证了企业设立激励机制对用户的创新意愿有着积极的影响[53]。赵婧（2021）利用小米社区数据以及问卷调查数据实证分析得出，越是想要得到社会认同感且具有较高的情感性态度，越愿意参与创新社区的[54]。刘伟和张波椋（2022）通过对小米MIUI社区的数据分析发现社区反馈的积极程度正向影响用户持续发表创新创意[55]。

企业在建立在线用户创新社区时，需要对用户进行合理的管理，不仅要对用户的创意动机掌握，还有对用户进行合理细分，因为一些价值用户提出的创意往往更具有实施价值。詹湘东（2013）认为在掌握用户参与创新动机的同时，应该也要注重对领先用户的识别，对创新社区的用户进行细分管理，为企业在获得有利价值创意提供便利[56]。Huang（2014）对戴尔创新社区的用户按照其潜在价值进行了合理细分，并且随着时间的推移发现低潜在价值用户活跃度会越来越低，而具有较高潜在价值的用户活跃度不变[57]。Brem和Bilgram（2015）研究发现，网络志和众包方法在社交媒体中可以有效地识别价值创新用户[58]。唐洪婷等（2017）通过超网络的理论将用户知识贡献联系集成超网络并建立模型，有效地识别了典型的在线用户创新社区领先用户[59]。王莉等（2019）认为社区影响力、活跃表现力、需求领先力为在线用户创新社区领先用户的三大特征，通过改进网络志的方法构建识别领先用户的模型，并验证了其有效性[60]。单晓红等（2021）将领先用户特征和传统个体特征进行融合，提出链路预测方法识别出华为创新社区不同类别的领先用户[61]。

### 2.2.2 创意价值影响因素相关研究

在创意理解方面，从其所包含的内容来看，创意具有四种不同维度：新颖性、可行性、相关性和详尽性。好的创意对企业改进产品和服务具有重要意义。在创意价值的影响因素方面；Hossain和Islam对Starbucks平台上被采纳的创意进行了分析，认为创意的价值很大程度上取决于其他用户的赞成票数和自身已发表的创意数[62]。Hossain和Islam从平台角度研究创意的价值，以Dall为例，在平台的早期成立阶段，创意实施比例较高，在平台发展成熟期，即使创意数量增长较快，实施比例依然较低。同时，创意的种类、投票数、是否具有参考创意也对创意价值具有影响[63]。Rauter通过文本分析技术研究了创意是否被价值与创意及其评论的文本内容之间的关系[64]。Li将创意特征与创意提交者特征作为影响因素，实证分析什么样的创意更具有价值，研究表明：创意提交者的参与经验以及创意价值比率、创意的受欢迎程度、长度、支持证据正向影响创意价值[65]；Schemmann研究对其他创意更加关注的用户所提交的创意、具有潜在创新性的创意以及受欢迎的创意更容易被采纳[66]。Wang采用精化似然模型来区分作为信息处理外围路线的快速和精益思想预选择过程，研究结果表明，用户地位、用户主动贡献和社区认同对理念预设有显著正向影响，而用户反应贡献则没有影响[67]。Wang认为大学产业协作（ UIC ）中日益增加的社区-创新现象提高成员企业的创新产出至关重要，并利用U-I共同资助数据，发现U-I创新社区成员适度的流动率能够提升成员企业的创新绩效，以及社区内部大学技术转移规模在社区动态性与事务所创新绩效之间起中介作用[68]。

企业建立在线用户创新社区不仅要促进平台用户持续贡献和分享优质创意，保持用户活跃性，还需要利用有限的人力和时间成本从大量的创意中筛选出有价值的创意，提高企业创新能力。如果企业没有办法及时了解用户需求，采纳合理的创意则会给企业带来负面影响，如果企业没有对用户的创意及时地提供反馈信息，正确处理用户创意，则会降低用户创新的积极性。郭琴基于星巴克的MyStarbucksIdea社区进行实证分析，结果发现创意受欢迎程度、创意复杂度、创意主观性、用户参与积极度、用户先前经验以及用户信誉显著影响创意实施价值，但在不同行业的开放式创新社区中，创意标题复杂度和创意内容复杂度对创意实施价值的影响存在一定的差异[69]。王婷婷（2018）研究了影响平台用户创意分享因素有哪些。结合前面用户行为特征社会资本的划分，对不同维度社会资本影响用户创意分享构建了模型[70]。李悦（2019）认为激励机制对用户创新动机有显著影响，通过构建创新价值影响因素假设模型与实证分析，验证了企业设立激励机制对用户贡献创新价值有着积极的影响[71]。祁运丽（2018）基于信息认知的理论视角，以小米在线社区MIUI论坛中的创意与用户为研究对象，实证分析探讨在线用户创新社区中创意特征与用户特征对创意采纳的影响，并分析社区已采纳创意、用户已发布创意的调节作用[72]。

### 2.2.3 创意选择方法相关研究

企业创建在线用户创新社区，吸收有效创意对自身产品创新，不仅要拥有合理的社区管理制度以及对价值用户的识别与区分，还要对创意的进行有效筛选和识别。就目前而言，企业对创意的识别主要还是通过成立有关部门进行人工筛选，但是人工筛选需要面临信息过载的巨大挑战，有时更会忽略具有价值意义的创意。因此，不少学者把研究角度转向创意选择方面。

Klein和Garcia（2015）认为传统的人工筛选方法存在主观性问题，提出了“柠檬袋”方法，通过对比分析证明其在快速筛选有价值创意方面的有效性[14]。Rhyn和Blohm（2017）采用文本挖掘和机器学习技术对创意的质量进行预测[73]。王婷婷（2018）研究了影响平台用户创意分享因素有哪些。结合前面用户行为特征社会资本的划分，对不同维度社会资本影响用户创意分享构建了模型[70]。祁运丽（2018）基于信息认知的理论视角，以小米在线社区MIUI论坛中的创意与用户为研究对象，实证分析探讨在线用户创新社区中创意特征与用户特征对创意采纳的影响，并分析社区已采纳创意、用户已发布创意的调节作用[74]。王松（2021）设计用户创意的双重网络结构，构建了基于图注意力网络的用户创意潜在价值发现模型，学习表达双重网络的节点特征及网络间映射关系，实现用户创意潜在价值早期发现[75]。马玉（2022）利用LDA主题模型对在线评论进行主题识别，提取出主要主题及每个主题下面的关键词。研究结果发现，在用LDA主题模型对在线评论进行主题识别时，可以较好地提取出文本数据中的关键要素[76]。薛水晶（2023）运用KNN、SVM、决策树、随机森林四种机器学习算法分析OpenIDEO在线用户创新社区数据，发现随机森林算法对创意的识别具有较高的准确性[77]。

### 2.2.4 研究述评

综上所述，在线用户创新社区的研究多集中在用户创新参与研究、用户细分及创新用户的识别、用户创意采纳研究等方面，但对创意价值影响机理及其筛选模型的构建还存在不足。

（１）通过对文献归纳与总结发现，在研究视角方面，创意特征与创客特征对创意价值都有着重要影响，但是现有研究分别以创意和创客两个视角单独研究，对创意与 创客双视角融合并没有得到足够的重视。在数据选取方面，现有研究多集中在创新社区现有的结构化的数据，对创意文本类型数据的研究甚少，本文认为 OIC 创意文本类型的诸多信息特征具有重要的分析价值。

（２）原有研究多是基于统计推断模型探究影响因素有哪些，虽然研究结论比较合理，但未从本质上解决人工筛选创意时效率低下的问题，因此本研究基于部分特征因素构建创意选择模型，帮助企业快速识别出具有价值的创意，及时了解用户兴趣趋向，减少用户等待反馈的时间，提高用户的参与感，为企业内部创意研究、开发提供可靠的指导。

（3）在当前学术研究中，对于创意选择的探讨相对较为有限，而且主要集中在机器学习现有的算法上对创意进行识别，缺乏对算法的进一步改进与优化的深入研究。在机器学习算法方面，有待进行更多的优化和创新，如结合集成学习、迁移学习等前沿技术以提高创意选择准确性。因此，为提高创意选择的准确性和全面性，有必要对现有算法进行进一步改进，应当集中在更准确地捕捉和评估创意的多维度特征，使算法更为适应复杂的创意表现形式。

## 2.3 本章小结

本章是论文的理论基础与支撑，围绕研究问题，回顾关于在线用户创新社区研究的相关理论和文献成果。关于理论基础，围绕开放式创新理论、在线用户创新社区概念、ELM模型、文本挖掘、Logistic回归模型、集成学习算法以及贝叶斯优化阐述和归纳，为本文研究夯实理论基础。关于文献综述，主要梳理和综述了在线用户创新社区、创意价值影响因素、创意选择和集成学习算法相关研究成果，为本文提供了丰富的文献支持。

# 第三章 研究假设与模型构建

通过对以往文献的整理可知，已有研究主要集中于可直接进行计量分析的创意和创意者的数值属性，如创意发布时间、创意者等级以及评论数量等。随着文本分析方法以及自然语言处理技术的发展，研究者开始探索创意文本内容相关属性变量对创意信息质量的影响，但以标题长度和创意极性为主，指标体系不够全面，对创意文本挖掘不够深入。除了结构化的特征之外，创意价值是否受到从创意文本的分析获得的特征的影响，以及如何影响，由ELM模型可知，信息内容和信息源特征共同影响信息的接受度。因此本文基于ELM模型将关于创意信息和创意者信息的众多影响因素沿着中心路径和边缘路径这两路径进行更加全面深刻的分析研究。其中中心路径假设包括创意文本类型特征和创意规范类型特征，边缘路径假设包括创客特征和用户评论特征。

## 3.1 中心路径假设

作为一种新兴的信息传播形式，创意以多种方式向用户和企业传递文本类型特征，其中包括标题、内容和情感倾向等。此外，创意还通过图片和点赞量等方式向用户和企业传递规范类型特征[78]。因此，在线用户创新社区创意价值的中心路径假设需要从以下与创意直接相关的几个方面进行深入分析：首先，需要考虑文本类型特征，包括文本详细度、文本可读性、文本相似度以及情感倾向（标题情感倾向、内容情感倾向）。其次，需要重视规范类型特征，包括创意含有的图片数量、创意得分数以及创意点赞数。这些特征能够有效地衡量创意的质量和受欢迎程度。

为了更好地理解创意的价值，需要深入探究这些特征之间的相互关系，以及它们与用户和企业之间的关系。通过对这些方面进行全面的分析，可以更好地评估和优化创意的质量和效果，从而为用户和企业带来更大的收益和价值。

### 3.1.1 创意文本类型特征

信息的完整性衡量的是信息发出者提供信息的全面性[79]。刘伟认为文本详细度高使读者更能清晰获取文本知识[80]，Wu在研究网络产品评论时发现，如果文本详细度不够，很难周密地将产品的全部特点表述清楚，而从详细度高的文本中可以获取到更有价值的信息[81]。本文认为，创客与企业之间的知识交互时，详细度不够的创意使企业难以识别出创意价值，企业为了能够快速筛选具有价值的创意，希望在识别创意时获取到更多的信息，由此提出以下假设。

H1：标题详细度对创意价值有正向影响。

H2：内容详细度对创意价值有正向影响。

在文本阅读理解方面“简单就是最好的”，更加复杂的结构，理解困难且说服力不强[82]，创意标题与内容的呈现是创客以文本形式对企业信息传递，Kara认为信息可读性高使企业付出较少的认知力度，可读性越好的文本可以清晰表达观点[83]，Chen指出阅读文本的容易程度是判断文本中是否含有价值信息的一个重要因素[84]。在OIC中，企业往往不会花费大量时间和精力去仔细推断文本含义，当创客要流畅地表达其创意的观点时，文本可读性越好越让企业以及社区用户更好理解创意作用，由此提出以下假设。

H3：标题可读性与创意价值有正向影响。

H4：内容可读性对创意价值有正向影响。

创意由创意标题和创意内容两个主题构成，而创意标题和创意内容之间的关系既独立又相关，创意标题是对创意内容的高度概括，信息相对简洁，在线用户创新社区主页面只显示创意标题，因此读者和企业首先会浏览到创意标题，当标题内容对其具 有吸引力的时候，他们更有可能继续阅读创意具体内容。简要且信息丰富的标题文本是吸引企业和其他用户浏览创意内容的一个重要前提。创意标题和创意内容是对创意以不同的表达方式来解释创意想法，因此两者具有一定的相似度。当二者相关性低时，则创意主题前后不一，企业将无法识别其价值所在，由此提出以下假设。

H5：标题与内容的文本相似度对创意价值有正向影响。

创意的文本内容主要是表达用户创意想法的文字，可能带有创意者的主观情绪和观点，包括正向的情感倾向，负向的情感倾向以及中性情感三种类型。当创客表达出对现有产品或者服务的抵触及批评情绪时会带有消极情感，而创意中表达了对自己创意和想法的肯定与自信情绪时会带有积极情感。说服性传播研究表示具有情感倾向的信息会影响信息说服的结果，虽然用户倾向于关注具有消极情感的信息，但是这类文本被认为可行性、可信度低，不具有客观性。研究发现情感会增加企业评估时的认知负荷，文本如果不具有明显的情感倾向，主要以中性客观的情感阐述相关信息，则该文本信息具有更高的有效性。在线用户创新社区中，创意标题与创意内容主要是创客向用户和企业阐述自己想法的文字表述方式，需要客观理性地描述创意，反而那些带有情感倾向的创意文本会影响其他用户和企业对创意质量的判断，创意主观性过高会让用户感觉创意表达不够正式，反而会影响到创意本身的价值，由此提出以下假设。

H6：标题情感倾向与创意价值呈倒U关系。

H7：内容情感倾向与创意价值呈倒U关系。

### 3.1.2 创意规范类型特征

当创意被发布后，其得分数往往成为衡量其受欢迎程度和质量的重要指标之一。在线用户创新社区设置机制来调整创意得分数的计算方式，以更准确地反映创意的质量和受欢迎程度，得分数计算的机制和权重设置决定了创意得分的真实性和公正性，这也保证了创意价值的客观性和可比性。这些得分数不仅反映了创意受欢迎程度和质量，而且还可能直接影响创意的传播和创新效果。同时，高得分的创意可能吸引更多用户关注，促进其进一步传播和营销，从而进一步提高其价值。创意得分数通过综合考虑创意的各项属性和用户评价，对创意的质量和受欢迎程度进行评价。创意的得分数在创意的认可、推广和价值提升方面具有重要的作用，是评估创意价值的重要标志之一。由此提出以下假设。

H8：创意得分数对创意价值有正向影响。

随着在线用户创新社区的普及和用户互动的增加，点赞数也是衡量创意价值和社交影响力的重要途径，点赞数越高的创意往往会获得更多的曝光和传播，吸引更多用户的关注和参与。这些用户可能会在自己的社交圈中分享和转发该创意，从而进一步扩大其影响力和传播范围。这种通过点赞数获取的社交效应有助于提升创意的社交影响力和传播效果，进而提高其价值。由此提出以下假设。

H9：创意点赞数对创意价值有正向影响。

## 3.2 边缘路径假设

与创意间接相关的特征，分别有创客特征和对创意评论的用户特征，作为创意的信息源，创客在企业创新发展中扮演着重要角色，他们通常具有对技术和科学的深入了解和熟练运用，能够从生活中发现和抓住创意的源头，并将其转化为具有实际价值的创意产品或服务。此外，扮演着对创意进行评价的角色，用户同样具有重要作用。因此，OIC创意价值的边缘路径假设需要从以下与创意间接相关的几个方面进行深入分析，首先，需要考虑创客特征，包括创客的创新能力、知识储备和先前经验。其次，需要重视用户特征，包括评论情感倾向。

### 3.2.1 创客特征

在线用户创新社区的旨在积极吸纳外部创新资源，以推动其内部的创新生态系统。在这一环境中，创客通过展现创意并获得相应的徽章数量，彰显其卓越的创新能力。这一徽章制度充分体现创客在创新社区中所作出的贡献，并提供一种客观的度量手段，以便准确评估和比较各位创客在创意产出方面的表现。在创新社区中，徽章数量不仅是对创客个体努力和创新贡献的一种公认的认可，同时也是评估其创意成果的一个关键指标。拥有更多徽章的创客通常意味着其在创新过程中表现出更高的水平和更为独到的见解，从而使其创意具有更高的价值。Siering等认为创新能力强的用户提供的知识资源更有作用[85]。Watts等研究在线平台时发现创意能力强的用户，其是更可靠的信息源[86]。作为创意提供者的创客，其获得的徽章数可能影响着创意价值，由此提出以下假设。

H10：创客的创意徽章数对创意价值有正向影响。

当创客在特定领域发布一定数量的创意时，将赋予其该领域的技能标签。技能标签被视为创客对特定领域的深刻理解和精湛掌握的象征，其数量反映了创客对多领域知识的广泛涵盖程度。拥有更多的技能标签表明创客具备更为丰富的知识体系和更广泛的专业涉猎，进一步显示其在多个领域的专业素养。辛冲技能标签数的增加不仅仅是数量的积累，更是对学科交叉和知识扩散程度的实质性反映[87]。因此，技能标签的数量在一定程度上反映了创客的学科交叉能力和知识扩散水平，进而影响其在创新领域的表现。技能标签的获取不仅是对个体创客努力的一种认可，同时也为整个创新社区的知识生态系统的发展提供了有力支持。由此提出以下假设。

H11：创客的技能标签数对创意价值有正向影响。

在创客提交创意的过程中，先前经验对不同方面都产生着深刻的影响。在获得徽章方面，先前经验的积累在提高创客对问题的敏感度和洞察力方面发挥着关键作用。通过先前的经验，创客能够更迅速地识别和理解需要解决的问题，从而提高创新的效率和效果。这种问题感知的敏感度使得创客能够更迅速地应对挑战，进而迅速积累徽章，作为对其创新能力的认可。在技能标签的层面，先前经验为个体提供了更好的理解和应用已有知识的能力。这使得创客能够更深入地掌握领域内的专业知识，进而更有针对性地为创新活动作出贡献。在创意的产生过程中，个体的知识储备是至关重要的基础，而先前经验有助于个体更深入地理解和运用这些知识，使其更容易发现并创造新的创意，由此提出以下假设。

H12a：创客先前经验正向调节创客的创意徽章数对创意价值的影响。

H12b：创客先前经验正向调节创客的技能标签数对创意价值的影响。

### 3.2.2 用户评论特征

用户评论在创意领域中具有重要的作用，它们是对创意的反馈和态度的直接体现。这种反馈包括对创意的赞赏、批评、建议等，为创意制作者提供了有价值的信息，帮助他们更好地了解用户需求和市场趋势。用户评论对于创意的优化和改进提供了有力的参考。其重要性体现在其能够直接影响其他用户的态度和行为。正面情感倾向的评论有助于提高创意的曝光率和认可度。这种积极的反馈可以吸引更多的用户和企业关注，促进创意的传播和推广。用户往往更愿意尝试被广受好评的创意，从而进一步增强创意在市场上的影响力。相反，负面情感倾向的评论可能对其他用户和企业对创意的信任和接受度产生负面影响。这种情况可能降低创意的影响力和市场竞争力，因为企业可能会受到他人的负面评价而产生犹豫或拒绝的态度。由此提出以下假设。

H13：用户评论情感倾向对创意价值有正向影响。

## 3.3 模型构建

为了控制一些不被观察到的因素影响创意实施价值，本文还从创意自身特征加入控制变量，包括创意提交时间和创客注册时间。结合研究假设，本文提出以下在线用户创新社区创意价值影响因素模型，见图3.1。

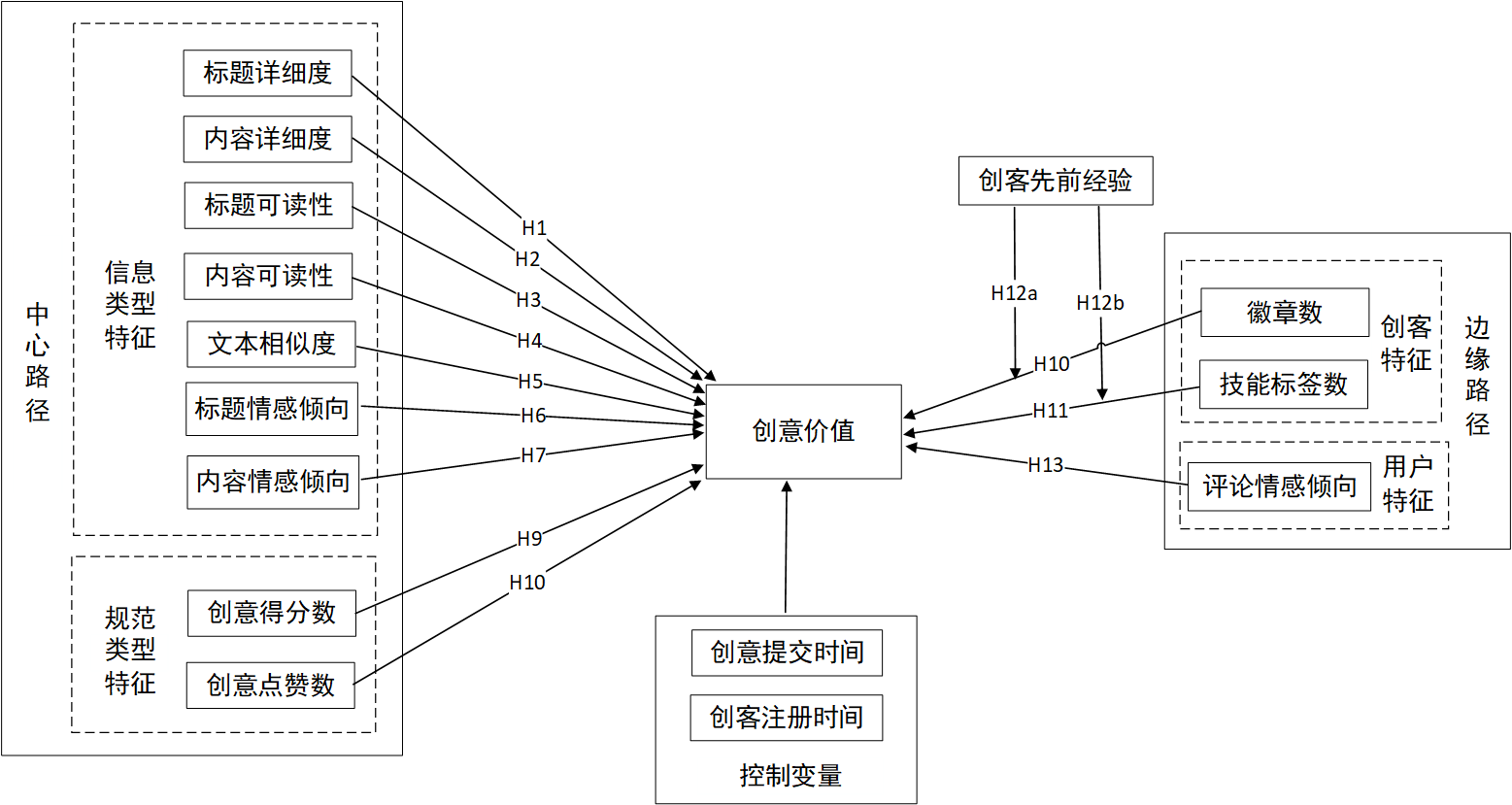


图 3.1 在线用户创新社区创意价值影响因素研究模型

Figure 3.1 Research model on the influencing factors of creative value in online user innovation communities

相关假设：

H1：标题详细度对创意价值有正向影响。

H2：内容详细度对创意价值有正向影响。

H3：标题可读性与创意价值有正向影响。

H4：内容可读性对创意价值有正向影响。

H5：标题与内容的文本相似度对创意价值有正向影响。

H6：标题情感倾向与创意价值呈倒U关系。

H7：内容情感倾向与创意价值呈倒U关系。

H8：创意得分数对创意价值有正向影响。

H9：创意点赞数对创意价值有正向影响。

H10：创客的创意徽章数对创意价值有正向影响。

H11：创客的技能标签数对创意价值有正向影响。

H12a：创客先前经验正向调节创客的创意徽章数对创意价值的影响。

H12b：创客先前经验正向调节创客的技能标签数对创意价值的影响。

H13：用户评论情感倾向对创意价值有正向影响。

## 3.4 本章小结

本章主要介绍了在线用户创新社区创意价值影响的理论假设，并提出了研究模型，对创意文本类型特征、创意规范类型特征、创客特征以及用户评论特征进行说明，旨在为第四章研究设计与实证分析奠定基础。

# 第四章 研究设计与实证分析

根据第三章的研究假设以及构建的在线用户创新社区创意价值影响因素研究模型，本章将在第三章的基础上进行实证分析，验证理论假设并深入了解各个因素在创意价值形成中的实质性影响。首先通过创新社区收集在线用户创新社区中的实际数据；接着对在线用户创新社区创意的文本特征，采用文本挖掘方法进行量化，并构建Logistic回归模型；最后对Logistic回归结果进行分析，为进一步在线用户创新社区创意选择模型的构建提供有效的支持。

## 4.1 研究对象与数据采集

本文的数据来源是Salesforce在线用户创新社区（社区链接：https://success.salesforce.com/ideaSearch?filter=All），社区界面如图4.1所示，该平台是由目前在OIC中最有影响力的大型客户关系管理的软件服务提供商。Salesforce在线用户创新社区公开所有创意相关信息和用户信息，可以详细地观察到创意从提交到采纳的具体状态以及用户之间的交互信息。网站具有良好的公开性，保证了创意数据和用户信息的可获得性。Salesforce为本文提供了足够的数据以供支撑，因此本文选取Salesforce研究对象。

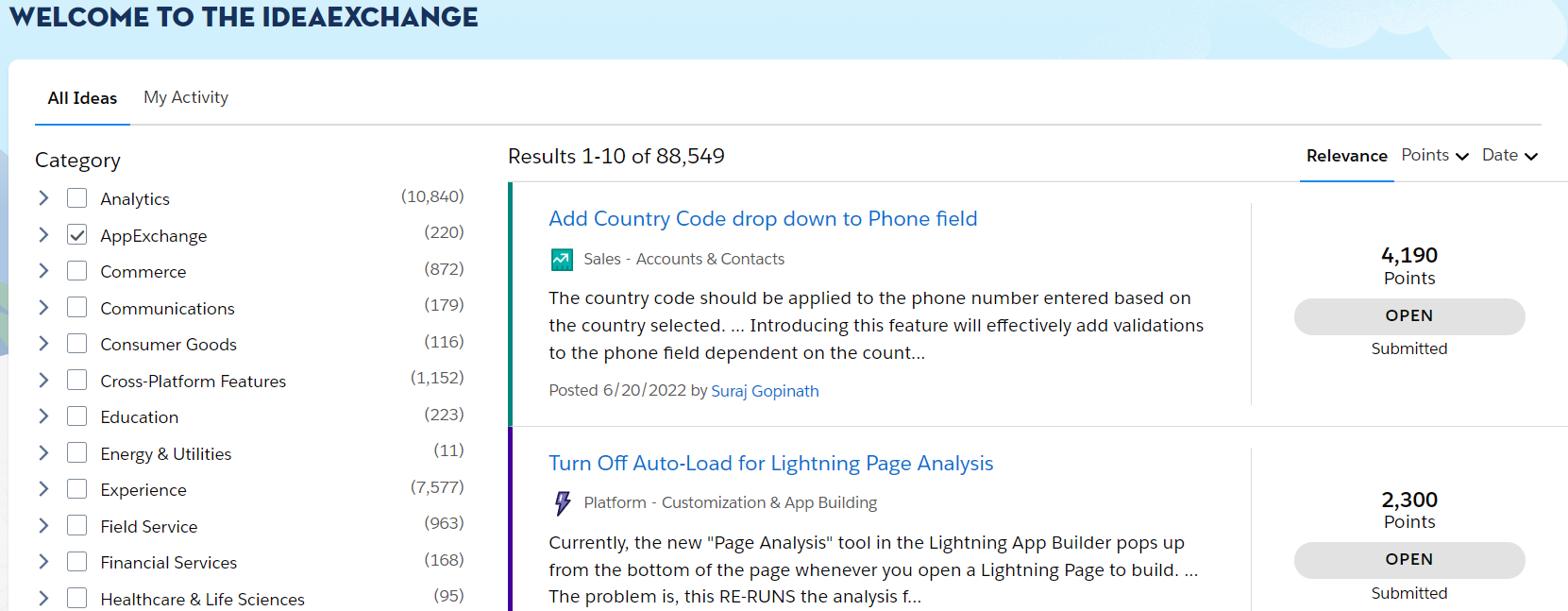


图 4.1 Salesforce在线用户创新社区界面（部分）

Figure 4.1 Salesforce Online User Innovation Community Interface (Partial)

利用Python自编数据采集程序进行数据采集，在中心路径假设中，采集的数据主要包括创意标题、创意内容以及图片数量，如图4.2所示。在边缘路径假设中，采集的数据主要包括创客的徽章数、技能标签数、创意总得分以及创意的评论内容，如图4.3所示。

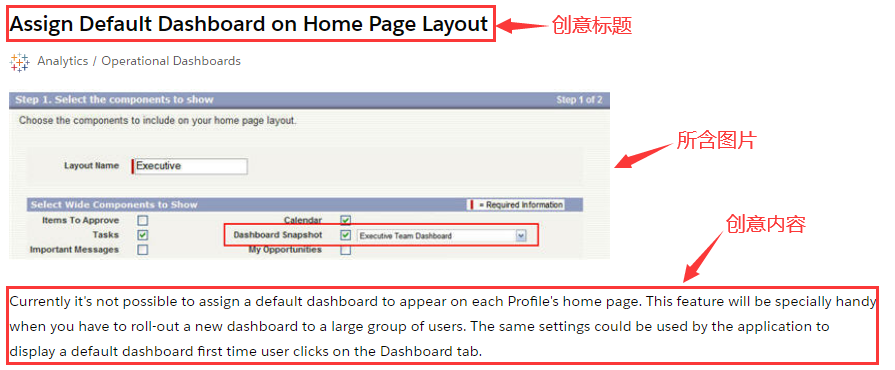


图 4.2 Salesforce 创意特征示例

Figure 4.2 Example of Salesforce Creative Features

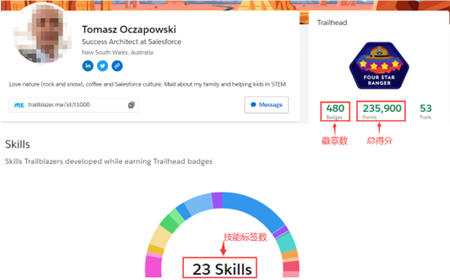
****

图 4.3 Salesforce 创客特征示例

Figure 4.3 Example of Salesforce Maker Features

本文于2022年8月对Salesforce在线用户创新社区数据进行采集，获取Analytics板块的CRM Analytics、operational Dashboards与Operational Reporting三个创意类型各项数据，经过清理、去重和筛选，最终获取到12558条有效创意。

## 4.2 变量获取

### 4.2.1 因变量

本文研究的是对创意价值的影响因素分析，因此将创意价值作为因变量，在 Salesforce 在线用户创新社区中，每个创意都会显示当前状态 ， 有 “ DELIVERED ”“ IN DEVELOPMENT”状态的创意代表具有创意价值，此状态用 1 表示，而有“OPEN”“Low Activity”等其他状态的创意代表不具有创意价值，此状态用 0 表示。

### 4.2.2 自变量

**（1）文本详细度**

本文通过分析单词复杂性和句子结构来进行计算。首先将文本分割成句子，并对每个句子的结构复杂性进行评估，包括句子长度和句式数量等。接着将句子拆分成单词，并评估每个单词的复杂性，如长度、词频和是否为专业术语。最后，通过对所有单词和句子的复杂性进行统计，计算平均值，并综合考虑字词复杂性和句子结构的贡献，得出文本详细度得分，文本详细度 = (Σ(单词复杂性) + Σ(句子结构复杂性)) / (总单词数 + 总句子数)。公式如（4.1）所示：

其中，为文本中的总单词数，为文本中的总句子数；和 是对应单词和句子的权重，用于平衡单词复杂性和句子结构复杂性的重要性。

**（2）文本可读性**

本文采用Flesch Reading Ease（FRE）公式计算文本可读性[88]，FRE公式是Rudolf Flesch教授创建用来计算文本可读性的一种标准公式，以提供一个量化的指标，帮助人们更好地理解和改善文本的可读性水平。它通过分析文本中的句子长度和单词音节数来计算一个分数。该分数越高，表示文本越容易理解。其评分范围为0到100，分数越高表示文本越容易理解。如公式（4.2）所示：

其中，为平均句子长度（即单词数除以句子数）；为每个单词的平均音节数（即音节数除以单词数）。

**（3）文本相似度**

本文采用Python编程程序实现文本相似性计算的方法。首对文本数据进行多步预处理以确保有效性，包括去除标点符号和数字，将所有单词转换为小写形式，以及执行分词和停用词去除操作。特别地，将创意标题和创意内容分别视为两个单词数组。对于创意内容，计算每个词语在其中的词频，并过滤掉低频词语，以确保只考虑在创意内容中具有足够频率的词汇。接下来，构建一个语料库词典，将每个文档表示为一个向量，其中每个元素是由词的编号和其在文档中的词频组成的二元组。这样的表示方式为后续的文本比较提供了基础。通过使用字典方法，获取了文档的词袋（bag-of-words），将对比文档加载并转换为相应的词袋模型。再进一步的处理中，利用gensim库中的tf-idf模块计算了每个单词的权重，为了更准确地反映每个词语在文本中的重要性，以便更有效地进行相似性度量。最终，使用余弦相似度作为度量标准，衡量每个创意标题及其相应创意内容之间的相似度。

余弦算法是NLP中一种原理简单、结果又易于观察的文本相似度算法[35]。它基于文本的向量表示，使得非结构化的文本能够更方便地进行数学计算和相似度比较。

给定创意标题文本向量A和创意内容文本B，二者余弦相似度的计算如公式（4.3）所示：

其中，表示向量内积，即对应维度上的元素相乘后求和。而和分别表示向量 A 和 B 的范数（或长度），计算方式为对应维度上的元素平方和后再开平方。

而这个相似度值的范围在 -1 到 1 之间，其中 1 表示完全相似，-1 表示完全不同，0 表示无相似性。在实际应用中，通常将余弦相似度值归一化到 [0, 1] 范围，通过公式（4.4）所示：

这样可以更方便地解释相似度的意义，使得 0 表示不相似，1 表示完全相似。

**（4）情感倾向**

情感分析是NLP中常见的应用，它是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。计算情感得分：对于每个文本，统计其中出现的情感词和它们的情感强度值，然后计算文本的情感得分。本文使用情感得分加权求和方法，将文本中所有情感词的情感强度值加权求和，得到文本的情感得分。计算方法如公式（3）所示：

其中，PosiWords和PassWords分别为正向情感单词数量和负向情感单词频率，PosiSenti和PassSenti分别为正向情感强度与负向情感强度。例如，假设文本包含两个情感词“happy”和“sad”，它们的情感强度分别为0.8和-0.6，出现的频率分别为2和1，则文本的情感得分为(0.8\*2 + (-0.6)\*1) / 3 = 0.33。

在基于Python编程的情感分析过程中，编写了Sentistr作为情感强度计算工具。该过程主要包括文本预处理和情感得分计算两个主要步骤。文本预处理阶段包括小写化、词干化、词形化、去除标点和停用词以及规范化等步骤，以保障后续情感得分的准确性。在识别各词项的情感词汇时，借助斯坦福大学的CoreNLP成熟的词库，为每个词项分配情感得分，以表征其情感强度。最终，通过计算文档情感强度，以每个创意文本为单位，得出所有单词情感得分的均值并进行归一化计算，从而获得该创意文档的情感强度得分。该得分在-1到1之间取值，情感得分越接近-1表示负向情感越强烈，越接近0表示情感越中性，越接近1表示正向情感越强烈。

**（5）创意得分数**

在本文中，创意得分数作为一项重要的结构化数据进行采集与分析。创意得分数是指针对每个创意进行的定量评估，该评估基于特定标准和指标体系，旨在客观地衡量创意的质量和创新程度。OIC平台采用了一套严谨而全面的评分体系，该体系涵盖了创意的原创性、实用性、可行性、可持续性等多个方面。通过这个评分体系将创意得分量化呈现给其他用户，在数据采集过程中，本文针对每个创意的得分数将其记录在结构化数据集中。为确保数据的可信度和准确性，进行了多次验证和交叉核实，同时采用了数据清洗和异常值处理等方法，以保证数据的可靠性和一致性。

**（6）创意点赞数**

在线用户创新社区中，创意点赞数是指其他用户或社区成员对于特定创意项目表示认可和支持的数量。点赞数在一定程度上反映了创意价值。在本文中，将创意点赞数视为另一项重要的结构化数据，旨在通过社交互动的方式了解创意的价值，为获得准确的创意点赞数，同样进行了数据的验证和清洗工作，以确保数据的准确性和完整性。

**（7）创客徽章数**

在在线用户创新社区中，创客作为参与者，通过平台上传自己的创意。当这些创意经过仔细审查并被平台采纳后，便实时地应用于相应领域或项目中。在此过程中，对于每个被采纳的创意，平台会授予创客一个特殊的徽章，以表彰其创新贡献和卓越的创意表现。创意徽章被视为一种认可和奖励的象征，它不仅仅是一个虚拟的荣誉标志，更是对创客创新能力的有效衡量指标。通过获取这些徽章，创客能够在社区内树立起良好的声誉，进而吸引更多关注、支持和合作机会。

**（8）创客技能标签数**

在获得创意徽章的同时，每当创客获得不同类型的创意徽章，就会被授予相应类型的技能标签，象征着其在特定领域或专业方向上的专业知识和能力。因此，创客所获得的技能标签数量成为衡量其知识储备广度的重要标志。随着技能标签数量的增加，创客的知识储备不断丰富与拓展。这表明创客具备跨领域的学科视野和专业知识，有能力在不同领域进行创新性的探索与贡献。丰富的知识储备为创客提供了解决复杂问题和应对多样挑战的优势，从而进一步促进其在在线用户创新社区中的创意产出和贡献。

### 4.2.3 调节变量

**（1）先前经验**

在线用户创新社区中，即使创客未获得创意徽章或技能标签，其每个创意也会被平台进行打分。这些分数反映了创客在创意产出方面的表现，高分意味着创客具备一定的创新经验。通过统计创客所有创意得分的总和，可以评估其在社区内的创新历程，将其视为创客的先前经验，创客的先前经验在很大程度上可以反映出其创意产出的累积和发展过程。

### 4.2.4 控制变量

在创客特征方面，创客的注册时间、对创意的提交时间都可能会影响创意价值，因此这些结构化的数据可以直接从平台中爬取获得，可以直接作为控制变量应用。

变量具体定义如表4.1所示。

表 4.1 模型变量与定义

4.1 Model variables and definitions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **变量类型** | **特征类别** | **变量名** | **变量指标** | **变量定义** |
| 因变量 |  | 创意价值 | Value | 创意状态、具有价值标签为1；不具有价值标签为0。 |
| 自变量 | 创意文本类型特征 | 标题详细度 | TitleDetail | 创意标题详细度，[0,100] |
| 内容详细度 | ConDetail | 创意内容详细度，[0,100] |
| 标题可读性 | TitleRa | 创意标题可读性，[0,100] |
| 内容可读性 | ConRa | 创意内容可读性，[0,100] |
| 文本相似度 | TextSimi | 创意标题与创意内容之间文本相似度，[0,100] |
| 标题情感倾向 | TitleSenti | 创意标题的情感得分，[-1,1] |
| 内容情感倾向 | ConSenti | 创意内容的情感得分，[-1,1] |
| 创意规范类型特征 | 创意得分数 | CreaPoint | 创意得分数 |
| 创意点赞数 | CreaLikes | 创意点赞数 |
| 创客特征 | 徽章数 | BadgesNum | 创客在在线用户创新社区获得的创意徽章数 |
| 技能标签数 | SkillsNum | 创客在在线用户创新社区获得的技能标签数 |
| 用户评论特征 | 评论情感倾向 | ComSenti | 用户对创意评论文本的情感得分，[-1,1] |
| 调节变量 |  | 先前经验 | PreExperience | 创客历史创意总得分表示创客的先前经验 |

## 4.3 描述性统计与相关性分析

### 4.3.1 描述性统计

表4.2为对各变量初始值的描述性统计，可以据此了解各变量的数值分布情况。

表 4.2 描述性统计分析

Table 4.2 Descriptive Statistical Analysis

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **变量名** | **样本量** | **最大值** | **最小值** | **平均值** | **标准差** | **中位数** |
| Value | 12558 | 1 | 0 | 0.242 | 0.342 | 0.03 |
| TitleDetail | 12558 | 50 | 1 | 32.269 | 10.673 | 33 |
| ConDetail | 12558 | 5260 | 20 | 2597.849 | 1357.434 | 2575 |
| TitleRa | 12558 | 99.998 | 1.084 | 66.84 | 22.192 | 70.913 |
| ConRa | 12558 | 99.981 | 1.063 | 60.247 | 20.902 | 60.684 |
| TextSimi | 12558 | 99.976 | 1.01 | 59.732 | 23.637 | 63.103 |
| TitleSenti | 12558 | 0.998 | -0.982 | 0.236 | 0.21 | 0.183 |
| ConSenti | 12558 | 0.999 | -0.992 | 0.296 | 0.259 | 0.23 |
| CreaPoint | 12558 | 13087 | 4 | 6407.534 | 3329.446 | 6326 |
| CreaLikes | 12558 | 7886 | 3 | 2733.279 | 1520.836 | 2659 |
| BadgesNum | 12558 | 35 | 1 | 16.688 | 9.051 | 16 |
| SkillsNum | 12558 | 27 | 1 | 14.415 | 6.993 | 15 |
| ComSenti | 12558 | 0.999 | -0.972 | -0.006 | 0.543 | -0.012 |
| PreExperience | 12558 | 39812 | 156 | 20235.756 | 10672.739 | 19338 |

统计得出，创意价值（Value）均值为0.242，即具有创意价值的创意为条3043，占创意总数仅为 24.3%，而不具有创意价值的创意为9515条，占创意总数为75.7%，说明价值创意与非价值创意的离散程度很高，在线用户创新社区具有创意价值的创意少之又少，因此分析创意价值的影响因素，通过影响因素作为创意选择的参考依据，存在分析价值。

### 4.3.2 相关性分析

相关性分析用于评估变量之间是否存在显著的相关性。皮尔逊相关系数是衡量变量关系强弱的指标，其正负符号表示正负相关。在相关性分析中，绝对值大于0.8的相关系数可能存在共线性问题，即变量之间存在高度线性关系，这可能导致模型的不稳定性和解释困难。通常，0.5 < |r| < 0.8被认为是中度相关，0.3 < |r| < 0.5为低度相关，而|r| < 0.3则表示弱相关。

在具体的研究中，通过观察表4.3中的皮尔逊相关系数，可以得知所有相关系数绝对值均低于0.5，这一结果表明，研究并未出现严重的共线性问题，即使存在一定程度的相关性，但并未达到影响模型稳定性的水平。因此，得以支持将所有变量纳入模型中，这为后续的研究提供了坚实的基础，使得本文能够更加可靠地分析和解释变量之间的关系，为创意价值的研究提供了有力支持。为了更直观地展示，本文通过python中的matplotlib.pyplot工具包绘制了变量相关性的热力图，如图4.4。

表 4.3 变量之间相关性

Table 4.3 Correlation between variables

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Value | TitleDetail | ConDetail | TitleRa | ConRa | TextSimi | TitleSenti |
| Value | 1 |  |  |  |  |  |  |
| TitleDetail | 0.174\*\*\* | 1 |  |  |  |  |  |
| ConDetail | 0.098 | 0.168 | 1 |  |  |  |  |
| TitleRa | 0.431\*\* | 0.115\*\*\* | 0.106 | 1 |  |  |  |
| ConRa | 0.18 | 0.159 | 0.074\*\*\* | 0.172 | 1 |  |  |
| TextSimi | 0.303 | 0.148\*\*\* | 0.06 | 0.096 | -0.003 | 1 |  |
| TitleSenti | -0.083\*\* | 0.112 | 0.089\*\*\* | 0.105 | -0.019 | 0.098 | 1 |
| ConSenti | -0.046 | -0.084\*\*\* | 0.064 | 0.142\*\*\* | 0.01 | 0.059 | 0.091 |
| CreaPoint | 0.3\*\*\* | -0.032 | -0.073\*\*\* | 0.104\*\* | 0.005 | 0.122\*\*\* | 0.064 |
| CreaLikes | 0.291 | 0.135 | -0.015 | -0.102\*\*\* | 0.021 | 0.103 | 0.156 |
| BadgesNum | 0.016 | 0.13\*\*\* | 0.061 | -0.031\*\*\* | 0.002\*\*\* | 0.11\*\*\* | 0.094 |
| SkillsNum | 0.15 | 0.035 | 0.059 | 0.076\*\*\* | -0.013 | -0.053\*\*\* | 0.163 |
| ComSenti | 0.099\*\*\* | 0.156 | -0.007 | 0.08 | 0.013 | -0.022\*\*\* | -0.025\*\*\* |
| PreExperience | 0.299\*\*\* | 0.095\*\*\* | 0.103\*\*\* | -0.008\*\*\* | 0.001 | 0.132 | -0.025 |
|  | ConSenti | CreaPoint | CreaLikes | BadgesNum | SkillsNum | ComSenti | PreExperience |
| ConSenti | 1 |  |  |  |  |  |  |
| CreaPoint | -0.033 | 1 |  |  |  |  |  |
| CreaLikes | -0.037\* | -0.092 | 1 |  |  |  |  |
| BadgesNum | -0.035 | -0.059\*\* | 0.114\*\*\* | 1 |  |  |  |
| SkillsNum | -0.035\*\* | -0.037 | 0.068\*\*\* | 0.161 | 1 |  |  |
| ComSenti | -0.037 | -0.059 | 0.153\*\*\* | 0.116\*\*\* | 0.09 | 1 |  |
| PreExperience | 0.017\*\*\* | -0.033 | 0.115 | 0.116 | 0.077\*\*\* | 0.133 | 1 |

**注：\*, \*\*, \*\*\*表示在10%、5%和1%的水平下显著**

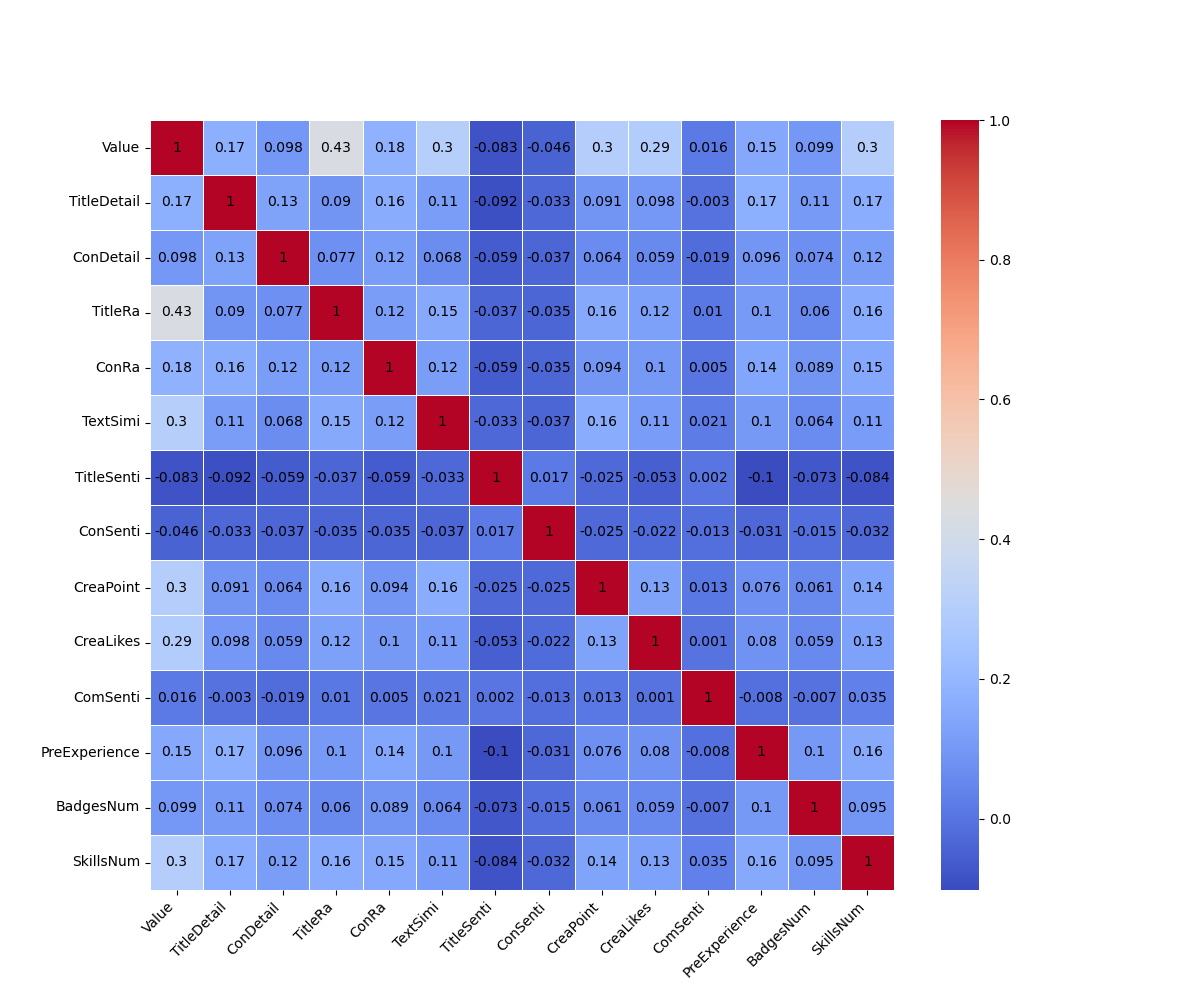
****

图 4.4 变量相关性热力图

Figure 4.4 Heat map of variable correlation

## 4.4 Logistic回归模型建立

本文创意价值的影响因素研究中，因变量创意价值是离散型变量，不适用于线性回归模型，对于一个二分类被解释变量的分析多使用非线性函数，最常用的是逻辑回归模型。逻辑回归是一种概率模型，其被解释变量为响应变量中某一结果发生与否的概率，自变量（包括解释变量和控制变量）为影响该结果发生的因素。根据上述变量测定和逻辑回归原理建立模型如下：

其中，P是CreaValue = 1的概率；1-P是CreaValue = 0的概率。由公式可以推出：

模型的自变量系数反映了创意有价值的概率在该自变量单位变化时的对数几率变化情况。在本文中，每个自变量的系数代表了与创意价值相关的特定影响因素对创意有价值概率的非线性影响。自变量系数的正负符号揭示了该影响因素对创意价值的影响正负，若为正，则表明该因素与创意价值呈正相关，即随着自变量的增加，创意有价值概率也相应增加；相反，若为负，则说明该因素与创意价值呈负相关，即随着自变量的增加，创意有价值概率会减少。此外，自变量系数的绝对值大小反映了影响因素对创意有价值概率变化的强弱程度。较大的系数表示该因素对创意价值的影响更为显著，而较小的系数则暗示着其影响较为微弱。因此，在解释模型结果时，应特别关注各自变量系数的正负符号和相对大小，以准确把握影响因素对创意价值的贡献程度。

## 4.5 Logistic回归结果分析

本文基于ELM模型进行影响因素分析，采用自变量分块进入模型，模型一的自变量选择边缘路径假设的创客特征和用户评论特征，包含以及徽章数、技能标签数、先前经验与徽章数交互项、先前经验与技能标签数交互性以及用户评论情感倾向；模型二在模型一的基础上加入中心路径假设的情感特征倾向，包含标题情感倾向和内容情感倾向；模型三在模型二的基础上将所有自变量纳入模型中。在提出的假设中，创意标题与创意内容具有的情感倾向与创意价值之间呈倒U型关系，因此在回归模型中需处理为平方值来验证。Logistic回归模型的三个模型的结果如表4.4所示，通过NagelKerke R²可以看出，三个模型的解释度是逐渐增加的，即加入新变量后模型的拟合度更优。

表 4.4 Logistic回归结果

4.4 Logistic regression results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 模型1 | P值 | 模型2 | P值 | 模型3 | P值 |
| TitleDetail |  |  |  |  | 0.330\*\*\* | 0.001 |
| ConDetail |  |  |  |  | 0.000 | 0.847 |
| TitleRa |  |  |  |  | 0.070\*\*\* | 0.001 |
| ContentRa |  |  |  |  | 0.014\*\*\* | 0.001 |
| TextSimi |  |  |  |  | 0.234\*\*\* | 0.001 |
| TitleSenti² |  |  |  |  | -0.065\*\*\* | 0.001 |
| ConSenti² |  |  |  |  | -0.347\*\*\* | 0.006 |
| Point |  |  | 0.243\*\*\* | 0.003 | 0.244\*\*\* | 0.002 |
| LikeCount |  |  | 0.021\*\*\* | 0.001 | 0.031\*\*\* | 0.004 |
| BadgesNum | 0.001\*\*\* | 0.001 | 0.002\*\*\* | 0.001 | 0.002\*\*\* | 0.001 |
| SkillNum | 0.095\*\*\* | 0.001 | 0.094\*\*\* | 0.001 | 0.092\*\*\* | 0.001 |
| SumPoint\* BadgesNum | 0.003\*\*\* | 0.005 | 0.001\*\*\* | 0.001 | 0.002\*\* | 0.06 |
| SumPoint\* SkillNum | 0.000 | 0.602 | 0.000 | 0.598 | 0.001 | 0.768 |
| ComSenti | 0.030 | 0.530 | 0.029 | 0.551 | 0.050 | 0.387 |
| NagelKerke R² | 0.552 |  | 0.653 |  | 0.932 |  |
| Hosmer-Leneshow | 0.816 |  | 0.823 |  | 0.833 |  |

注： \*, \*\*, \*\*\*分别表示在10%、5%和1%水平下显著。

上述模型中各自变量对创意价值影响具体分析结果如下：

在中心路径假设中，本文发现标题详细度对创意价值产生正向影响，较为详细的创意标题可以更清晰地传达创意的主要内容和特点，帮助企业或用户更好地理解创意，从而增加创意的价值，支持假设1。然而，研究结果表明，内容详细度并不显著影响创意价值，究其原因，过于详细的文本会将大量信息呈现在读者面前，会导致读者对文本信息感到混乱和压抑，难以捕捉信息的关键点和主旨，不支持假设2。标题的可读性和内容的可读性对创意价值产生正向影响，即可读性高的标题更能吸引用户的注意力，并且可以更快地让企业和用户理解创意的主要内容，从而有助于提升创意的价值。可读性高的内容使企业和用户更轻松地理解创意，提高创意被接受和传播的可能性，因此对创意价值产生积极影响，支持假设3和假设4。此外，标题和内容的文本相似度也对创意价值产生正向影响，即标题和内容相似的创意可以保持一致的信息呈现，增强用户对创意的理解，从而有助于提升创意的价值，支持假设5。另外，标题和内容的情感倾向的平方值对创意价值产生负向影响，这表明带有主观情感的创意不会得到用户或企业的认可，而越客观的创意越具有创意价值，即创意标题和内容的情感倾向与创意价值呈倒U型关系，支持假设6和假设7。除此之外，研究结果显示，创意得分数对创意价值产生正向影响，即当一个创意得分较高时，通常意味着该创意在创新性和解决问题上表现出色，这使得它更有可能为企业或用户带来实际价值。创意得分越高，创意价值越大，支持假设8。创意点赞数也对创意价值有正向影响，更多的点赞意味着创意更受欢迎，从而提高创意的价值，支持假设9。

在边缘路径中，根据研究结果，创客的徽章数正向影响创意价值，创客在在线用户创新社区中获得越多徽章表明其在社区中具有更高的影响力和创新能力，从而使其创意更受关注，提升创意的价值，支持假设10。创客的技能标签数正向影响创意价值，创客拥有更多的技能标签意味着其在多个领域拥有专业知识，这可以使其创意更具多样性和实用性，从而增加了创意的价值。不同领域的知识和技能可以为创意提供更丰富的背景和视角，有助于创意更好地解决现实问题或满足用户需求，支持假设11。创意获得的徽章数与创意总得分的交互作用会进一步强化创客的获得的徽章数与创意价值之间的正向关系，支持假设12a。技能标签数与创意总得分的交互作用未显示对创意价值的显著影响，即创客的总得分并不正向调节创客的技能标签数对创意价值的影响，不支持假设12b。用户评论的情感倾向对创意价值的影响不显著，创意的价值主要受创意本身的特点和质量所影响。虽然评论可以提供反馈和洞察，但创意的独创性、实用性以及解决问题的能力更可能是决定创意价值的关键因素，不支持假设13。

通过Logistic回归结果分析，研究假设检验结果汇总如表 4.5 所示：

表 4.5 研究假设结果汇总

Table 4.5 Summary of research hypothesis results

|  |  |
| --- | --- |
| 研究假设 | 结果 |
| H1：标题详细度对创意价值有正向影响。 | 支持 |
| H2：内容详细度对创意价值有正向影响。 | 不支持 |
| H3：标题可读性与创意价值有正向影响。 | 支持 |
| H4：内容可读性对创意价值有正向影响。 | 支持 |
| H5：标题与内容的文本相似度对创意价值有正向影响。 | 支持 |
| H6：标题情感倾向与创意价值呈倒U关系。 | 支持 |
| H7：内容情感倾向与创意价值呈倒U关系。 | 支持 |
| H8：创意得分数对创意价值有正向影响。 | 支持 |
| H9：创意点赞数对创意价值有正向影响。 | 支持 |
| H10：创客的创意徽章数对创意价值有正向影响。 | 支持 |
| H11：创客的技能标签数对创意价值有正向影响。 | 支持 |
| H12a：创客先前经验正向调节创客的创意徽章数对创意价值的影响。 | 支持 |
| H12b：创客先前经验正向调节创客的技能标签数对创意价值的影响。 | 不支持 |
| H13：用户评论情感倾向对创意价值有正向影响。 | 不支持 |

## 4.6 本章小结

本章主要描述了数据处理与变量获取过程，数据处理部分通过NLP技术对文本类型数据进行量化的过程，主要包括文本详细度、文本可读性、文本相似度以及文本的情感倾向，并且介绍了规范类型数据的过去过程，然后对模型进行Logistic回归分析并得出Logistic回归分析结果，获得了在线用户创新社区显著影响创意价值的影响因素。

# 第五章 在线用户创新社区创意选择方法研究

第四章中已对在线用户创新社区创意价值的影响因素进行了研究分析，证明本文提出的大部分创意和创客特征对创意价值具有显著的影响。但是研究创意价值的影响因素实现创意自动筛选仍然困难。以往研究证明基于文本挖掘和机器学习算法的预测方法是处理在线用户创新社区中大量创意的可行方法，以快速、可重复的方式从数据中发现模式和提取有用信息，因此本章基于这些具有显著性的影响因素，结合分类预测技术构建在线用户创新社区价值创意选择模型，建立自动化创意筛选系统，为企业实现高效率地创意筛选提供指导意见。

## 5.1 创意选择模型框架

在线用户创新社区创意选择模型总体框架流程图如图5.1所示。基于第四章影响因素分析结果，数据样本包含多个关键特征，包括创意文本类型特征、创意规范类型特征和创客特征，数据样本将被分为训练集和测试集两个部分输入到选择模型中，然后得到相应的选择结果。然而，单一分类器的性能评估不能保证在整个数据集上实现最佳预测性能，为了最大程度地充分利用实验数据，并确保所构建的预测模型在实际应用中表现出色，本文选取多种流行的集成学习分类算法，将在这些算法中选择最适合的，并对其进行进一步优化根据输出具有价值的创意的概率值，在此基础上建立在线用户创新社区创意选择模型。

本文将从Saleforces在线用户创新社区中获取的创意数据进行一系列预处理过程，进而获得创意和创客文本类型特征以及规范类型特征等实验数据，然后结合RF（随机森林）、GBDT（梯度提升决策树）、AdaBoost、XGBoost以及LightGBM这五种集成学习算法对创意价值进行预测。首先利用训练数据集训练预测模型，然后将测试集输入到训练好的预测模型中，通过得到的预测结果选择效果突出的模型并进行优化，最后再利用新的数据集对在线用户创新社区创意价值进行预测。

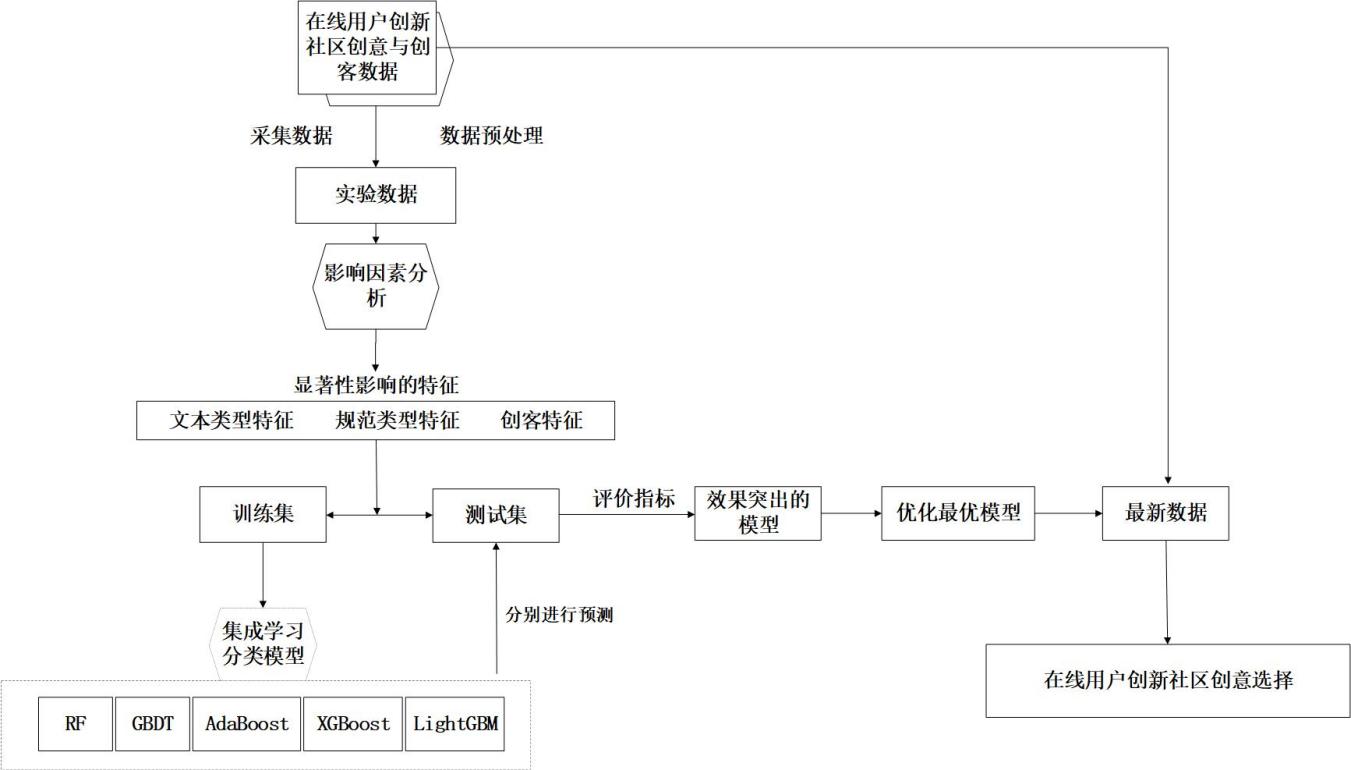


图 5.1 在线用户创新社区创意选择模型框架

Figure 5.1 Framework of Creative Selection Model for Online User Innovation Community

## 5.2 集成分类器的评价指标与选择

### 5.2.1 评价指标

本文的目标是对具有价值的创意进行选择，即创意具有价值或不具有价值，因此转化为一个分类问题，以便使用机器学习方法解决，在分类模型评估中，基本的工具是混淆矩阵，它提供了关于模型分类性能的详细信息。对于二分类问题，混淆矩阵包括四个重要的度量指标：True Positive（将正类预测为正类的个数，TP）、False Positive（将负类预测为正类的个数，FP）、False Negative（将正类预测为负类的个数，FN）、True Negative（将负类预测为负类的个数，TN）。指标提供了对模型在正类和负类的分类性能的全面评估，允许我们深入了解模型的优势和局限，如表5.1 所示。

表 5.1 混淆矩阵

Table 5.1 Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | | 创意选择类别 | |
| 价值 | 非价值 |
| 实际类别 | 价值 | True Positive  （TP） | False Positive  （FP） |
| 非价值 | True Negative  （TN） | True Negative  （TN） |

在混淆矩阵的基础上，二分类问题的选择效果通常利用精确度（Precision）、召回率（Recall）、F1-Score（F1）结合ROC曲线进行评估。

精确度是模型选择具有价值的创意中，实际也具有价值的创意的比例，它用于衡量模型的整体分类准确性，公式如（5.1）表示：

召回率指在实际具有价值的创意集合中，被选择出来具有价值的创意的比例，可用公式（5-2）表示：

F1是精确度（Precision）和召回率（Recall）的调和平均，用于综合考虑模型的精确性和完整性，它对不平衡数据集有较好的适用性，可以用公式（5.3）表示：

ROC曲线是一种以TPR（True Positive Rate，真正例率）为纵坐标，以FPR（False Positive Rate，假正例率）为横坐标绘制的曲线。它用于描述模型在不同阈值下的灵敏度（能够正确识别正样本的能力）和特异性（避免将负样本错误识别为正样本的能力）之间的相互关系。另外，AUC是ROC曲线下的面积，通常介于0到1之间。AUC提供了一个单一的数值，用于度量模型的精确度。AUC越接近1，表示模型的性能越好，意味着模型在各种不同阈值下都能较好地区分正样本和负样本。AUC的高值表明模型在不同情况下都能保持较高的灵敏度和特异性，而不受阈值选择的影响。

### 5.2.2 集成分类器的选择

Bagging和Boosting是常见的集成学习方法，它们都通过构建多个弱学习器并将它们组合成一个强学习器来提高分类或回归的性能。Bagging算法采用随机自主采样来构建多个弱学习器，通过投票或平均等方式将它们集成为一个强学习器。随机森林（RF）和极端随机树（ET）是常用的Bagging算法，在分类决策中被广泛应用。这些算法都能有效地提高模型的泛化能力和预测精度。Boosting算法通过反复训练一个模型并重点关注被错误分类的数据样本来构建一组弱学习器，以此逐步提高整个模型的性能。梯度提升决策树（GBDT）、极端梯度提升（XGBoost）、AdaBoost和轻量级GBM梯度提升机（LightGBM）是Boosting算法中常用的方法，它们都在数据挖掘比赛和实际应用中都有着重要作用。因此本文选取RF、ET、GBDT、XGBoost、AdaBoost和LightGBM五类集成学习方法进行单独预测，对比其预测效果，筛选出预测效果突出的集成学习算法。

本文着重于采用Python编程语言，因其丰富的内置软件包，为所使用的机器学习算法提供了接口。在机器学习中，通常要求训练数据集、测试数据集以及最终应用模型的新数据集相互独立，以避免将训练时的误差率引入到模型性能的评估中。鉴于实验数据相对有限，特别是相对于在线用户创新社区上的数据而言，采用了10折交叉验证方法，以确保结果的稳健性，集成分类器在线用户创新社区创意选择效果如表5-2所示，ROC曲线如图5.2所示。

表5.2 集成分类器在线用户创新社区创意选择效果

Table 5.2 The Creative Selection Effect of Integrated Classifier in Online User Innovation Community

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | F1值 | Recall | AUC |
| RF | 0.83187 | 0.90636 | 0.93550 | 0.92656 |
| GBDT | 0.87918 | 0.89283 | 0.90691 | 0.92428 |
| XGBoost | 0.88335 | 0.91220 | 0.94799 | 0.94725 |
| AdaBoost | 0.83040 | 0.90424 | 0.94249 | 0.93805 |
| LightGBM | 0.80559 | 0.89053 | 0.92550 | 0.91753 |

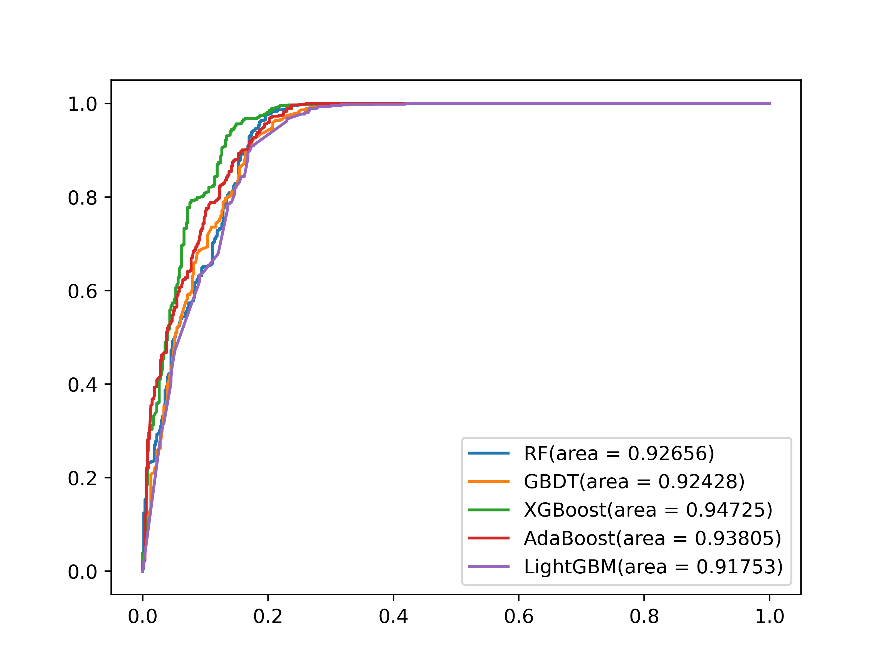


图 5.2 集成学习模型的ROC曲线图

Figure 5.2 ROC curve of ensemble learning model

观察到在Precision方面，XGBoost算法实现了0.88335的分数，与其他算法相比表现出色，表明XGBoost在预测正例时更具准确性，从而有望降低误报率。其次，F1值是Precision和Recall的综合指标，用于综合考虑模型的准确性和召回率，XGBoost在F1值方面达到了0.91220，意味着该算法在平衡准确性和召回率方面的表现较好，表明XGBoost在正类别的识别和误报控制方面更具竞争力。进一步观察到，XGBoost在Recall方面取得了0.98799的高分，说明该算法能够有效地捕获正类别，最大程度地减小漏报率。

最后， XGBoost在AUC方面实现了0.94725的分数，相对于其他算法，它表现出更高的曲线下面积，说明该算法具有更好的分类能力和更强的区分度。通过综合考虑Precision、F1值、Recall和AUC等性能指标，XGBoost在本文中表现出色。其高准确性、平衡的F1值、出色的召回率和卓越的AUC值，使其成为进一步深入研究和应用的有力选择。因此，本文选择XGBoost算法作为在线用户创新社区创意选择的优化算法。

## 5.3 改进XGBoost算法

XGBoost在这一领域展现出了卓越的性能。它以其出色的集成学习能力和梯度增强方法在分类问题中脱颖而出。然而，尽管XGBoost的卓越性能，仍有一些方面可以进一步优化，以提高其适应性和鲁棒性。

在数据集方面，本文注意到在线用户创新社区的数据往往存在不平衡性，即具有价值的创意远少于非价值的创意。为了解决这个问题，计划引入SMOTE（Synthetic Minority Over-sampling Technique）方法解决样本不均衡问题，通过合成少数类别的样本，来增加训练数据中的多数类别，从而平衡不平衡数据集。这将有助于提高XGBoost模型对于少数类别的识别和分类性能，使其更适合在线用户创新社区的实际应用。

在算法参数方面，将进一步改进XGBoost的超参数。尽管XGBoost以其可调参数众多而自由度高的特点而闻名，但参数的选择和调优仍然是一个挑战性任务。本文计划采用贝叶斯优化调参方法，来探索最佳的超参数组合。通过这种方式，可以精细调整XGBoost的性能，以更好地适应在线用户创新社区中的创意选择需求，具体优化方法如图5.3。

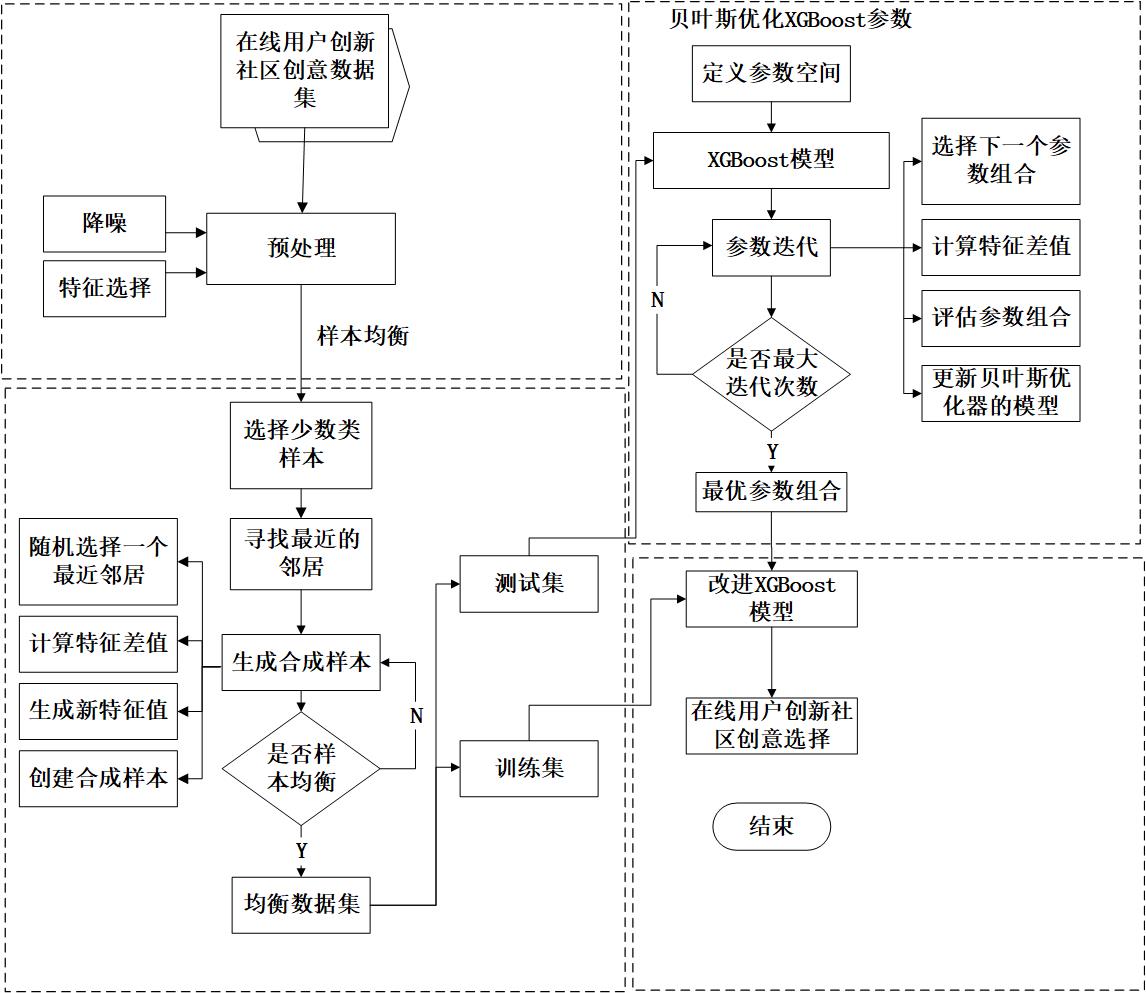


图 5.3 改进XGBoost算法过程图

Figure 5.3 Process diagram of improving XGBoost algorithm

### 5.3.1 基于SMOTE优化数据集

本文研究中，发现样本数据呈现显著的不均衡情况，其中创意采纳样本仅占总体样本的23%。这种不均衡现象在预测模型中产生了明显的负面影响，因为许多机器学习算法基于对数据的均匀分布进行假设，而实际数据却往往呈现不均衡的特征，这一不均衡导致预测模型更倾向于学习多数类样本的特征和模式，而相对忽视了少数类样本，从而损害了对少数类样本的准确预测能力，最终导致结果的明显偏差。

SMOTE算法，是21世纪初由学者Chawla提出的对少数类样本进行过采样的算法。其基本思想为，通过对样本的数据类别进行分析，确定少数类需要进行过采样的数量，然后按照相关规则重新生成新少数类样本，并与初始样本数据进行整合得到新研究样本，从而解决研究样本数据类别结构不均衡的问题，提高模型学习和泛化的能力。

本文采用SMOTE技术对少数类样本进行过采样，即通过SMOTE技术使具有价值创意与非价值创意进行样本均衡，原理如下：

1. 对于每一个具有创意价值的样本X，选择其k近邻的样本。
2. 选择一个过采样倍率N，以确定需要生成多少合成样本。通常，N是一个正整数。
3. 对于样本X，按照以下步骤生成合成样本：
4. 从其k个最近邻中随机选择一个样本，记作。
5. 对于每个特征j，计算合成样本特征：

其中，是一个从 [0, 1]之间随机选择的值，控制了新样本与原样本之间的差异程度。

重复步骤（3）直到生成N个合成样本，将生成的合成样本与原始数据集合并，以创建一个均衡的数据集。

### 5.3.2 基于贝叶斯方法优化参数

贝叶斯优化的目标是在有限的迭代次数内找到黑盒函数的全局最优解，无需了解函数的解析形式或求导。在XGBoost集成学习算法中，黑盒函数为模型验证集上的损失函数。贝叶斯优化将损失函数先验概率分布与观测到的数据（模型性能评估结果）结合起来，以获得后验概率分布，即对函数的最新估计。

对于贝叶斯方法改进XGBoost超参数流程如下：

1. 参数空间定义

本文选取了XGBoost集成学习算法中的四个关键参数，包括n\_estimators、

learning\_rate、max\_depth与subsample。n\_estimators表示迭代的次数，即基本决策树的数量。增加迭代次数可以提高模型性能，但也可能增加过拟合的风险；learning\_rate表示学习率，控制每次迭代中模型权重的更新幅度。较小的学习率通常需要更多的迭代次数，但有助于避免过拟合；max\_depth表示决策树的最大深度。较大的深度允许模型拟合更复杂的关系，但也容易导致过拟合；subsample训练每棵树时抽取的样本比例，用于防止过拟合。通常设置为小于1的值。具体参数空间范围如表5.3。

表 5.3 XGBoost集成学习算法参数空间定义

Table 5.3 Definition of parameter space for XGBoost ensemble learning algorithm

|  |  |
| --- | --- |
| 参数 | 范围 |
| n\_estimators | [50, 1000] |
| learning\_rate | [0.01,0.03] |
| max\_depth | [1, 100] |
| subsample | (0, 1] |

1. 目标函数定义

为了确保在参数搜索过程中正确地度量模型性能。在XGBoost模型的参数调优中，目标函数定义为评估模型性能的指标，函数将指导参数搜索以找到最佳的参数组合，从而最大化所选性能指标。在分类问题中，XGBoost的常见的性能指标包括Accuracy值和AUC值等。本文选择AUC值作为参数搜索过程中度量模型性能的指标。

1. 初始参数组合

为了避免模型过拟合问题，本文在验证集上进行交叉验证。在第一次迭代中，定义一组参数值，本文选取参数范围的中间值，目的在优化过程中加快迭代时间，参数值如表5.4所示。

表 5.4 贝叶斯优化参数初始值

Table 5.4 Initial values of Bayesian optimization parameters

|  |  |
| --- | --- |
| 参数 | 值 |
| n\_estimators | 275 |
| learning\_rate | 0.02 |
| max\_depth | 50 |
| subsample | 0.86 |

（4）进行迭代优化

在每次迭代中，将模型的性能和参数配置传递给贝叶斯优化算法，其会选择一组新的参数进行评估。根据评估结果更新模型性能结果，以便它可以生成下一组参数配置。本文使用Python编程语言中的BayesianOptimization工具包编写贝叶斯优化脚本，进行循环迭代，直到AUC值达到最优结束迭代。参数优化效果如图5.4所示，伪代码如表5.5所示。

表 5.5 贝叶斯改进XGBoost算法伪代码

Table 5.5 Pseudocode of Bayesian Improved XGBoost Algorithm

|  |
| --- |
| **伪代码** |
| data = 从 Excel 文件中读取数据('xgboost.xlsx') |
| X\_train, X\_test, y\_train, y\_test = 随机划分训练集和测试集(X, y, 测试集占比=0.2, 随机种子=42) |
| def 评估\_XGBoost(学习率, 最大深度, 树的数量, 子样本比例):  参数 = {  'learning\_rate': 学习率,  'max\_depth': 转为整数(最大深度),  'n\_estimators': 转为整数(树的数量),  'subsample': 子样本比例  }  return 准确率 + AUC |
| def 优化\_XGBoost():  贝叶斯优化器 = 创建 BayesianOptimization 对象(  目标函数=评估\_XGBoost,  参数边界={'learning\_rate': (0.01, 0.3),  'max\_depth': (1, 100),  'n\_estimators': (100, 1000),  'subsample': (0.1, 1)},  随机种子=42,  )  在初始化点上进行优化(初始点数=5, 迭代次数=10)  最佳参数 = 获取优化器的最大结果中的参数  return 最佳参数 |
| 输出 最佳参数 |

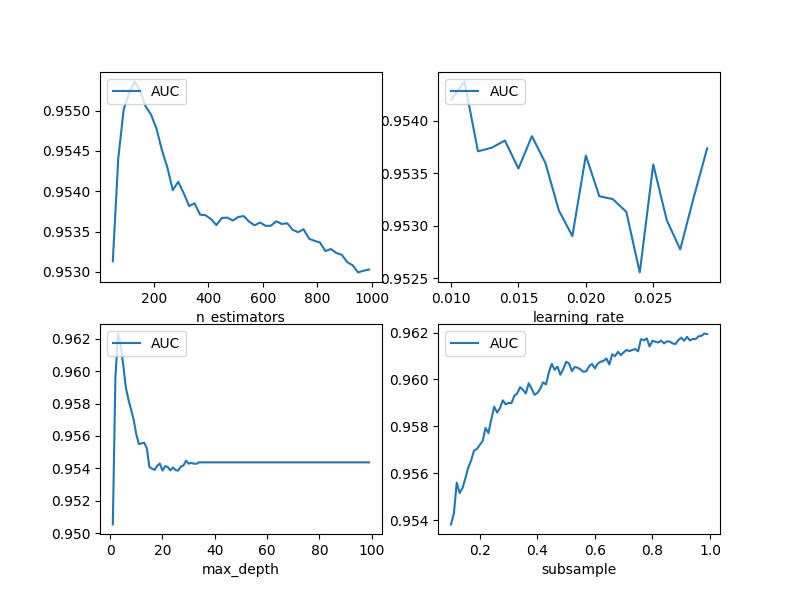


图 5.4 XGBoost参数优化效果

Figure 5.4 XGBoost parameter optimization effect

（5）确定XGBoost最优参数组合

在进行迭代优化的过程中，根据贝叶斯优化算法，逐步调整XGBoost模型的关键参数以最大化AUC值，最终得到的最优参数组合分别为n\_estimators为180、learning\_rate为0.011、max\_depth为4以及subsample为0.98，如表5.6。

表 5.6 XGBoost对在线用户创新社区创意选择的最优参数

Table 5.6 XGBoost's Optimal Parameters for Creative Choices in Online User Innovation Communities

|  |  |
| --- | --- |
| 参数 | 值 |
| n\_estimators | 180 |
| learning\_rate | 0.011 |
| max\_depth | 4 |
| subsample | 0.98 |

## 5.4 在线用户创新社区创意选择结果

根据上节优化方法论述，本文将运用SMOTE样本均衡方法对数据集进行调整，同时通过调整参数以获取优化后的XGBoost集成学习算法。在样本划分方面，按照8:2的比例将数据集分为训练集和测试集，旨在评估对在线用户创新社区的创意进行选择的效果。将所得结果与XGBoost集成学习算法进行对比，结果如表5.7所示。

表 5.7 改进XGboost模型与XGBoost模型的创意选择效果对比

Table 5.7 Comparison of creative selection effects between optimized XGboost model and XGBoost model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | F1值 | Recall | AUC |
| XGBoost | 0.88335 | 0.91220 | 0.94799 | 0.94725 |
| 改进XGBoost | 0.92381 | 0.93238 | 0.94892 | 0.96046 |

对比基本的XGBoost集成学习算法，改进后的模型在多个评估指标上都取得了显著的提升。精确度提高了约4.58%，F1值增加了约2.22%，召回率微幅提升约0.10%， AUC值则有着显著的1.39%提高。表明所采用的改进方法在样本均衡和参数调整方面取得了积极效果，有效地提高了在线用户创新社区创意选择的准确性和性能。

此外，为了更全面地评估模型性能，本文绘制了ROC曲线，并计算了相应的面积（AUC）。通过观察ROC曲线，清晰地看到优化的XGBoost集成学习算法的ROC曲线在整个面积下的AUC值为0.96046，相较于基本XGBoost的AUC值（0.94725）有显著提高。直观地反映了改进的XGBoost集成学习算法在捕捉正负例之间的平衡和对不同阈值的鲁棒性上的优势，ROC曲线图如图5.5所示。

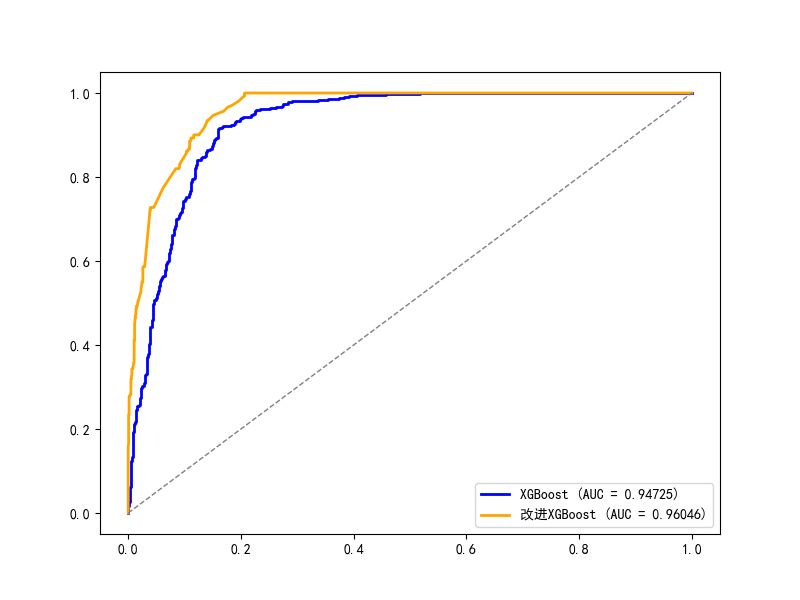


图 5.5 优化的XGboost模型与XGBoost模型的ROC曲线对比

Figure 5.5 Comparison of ROC curves between optimized XGboost model and XGBoost model

## 5.5 特征重要性分析

为了帮助企业更好地理解模型对于创意选择的影响，清晰地定位创意的重要特征。本文通过特征重要性分析，对在线用户创新社区创意选择所需要的特征进行重要性比较，分别计算特征在XGBoost集成学习算法的Mean Decrease Gini取值，其取值越大，说明特征越重要。结果如表5.8所示。

表 5.8 改进XGBoost集成学习算法后的特征重要性结果

Table 5.8 Results of feature importance after improving XGBoost ensemble learning algorithm

|  |  |  |
| --- | --- | --- |
| 特征 | 特征名称 | 特征重要性 |
| TitleRa | 标题可读性 | 0.73423 |
| TextSimi | 文本相似度 | 0. 61201 |
| TitleSenti | 标题情感倾向 | 0.34920 |
| ConRa | 内容可读性 | 0.27814 |
| ConSenti | 内容情感倾向 | 0.23989 |
| TitleDetail | 标题详细度 | 0.21119 |
| TitleSenti | 创意得分数 | 0.19837 |
| SumPoint\* SumBadges | 总得分×徽章数 | 0.11476 |
| CreaLikes | 创意点赞数 | 0.03036 |
| Knowledge | 技能标签数 | 0.01086 |

为使得结果更直观，对上表进行统计和可视化分析整理后如图5.6所示。

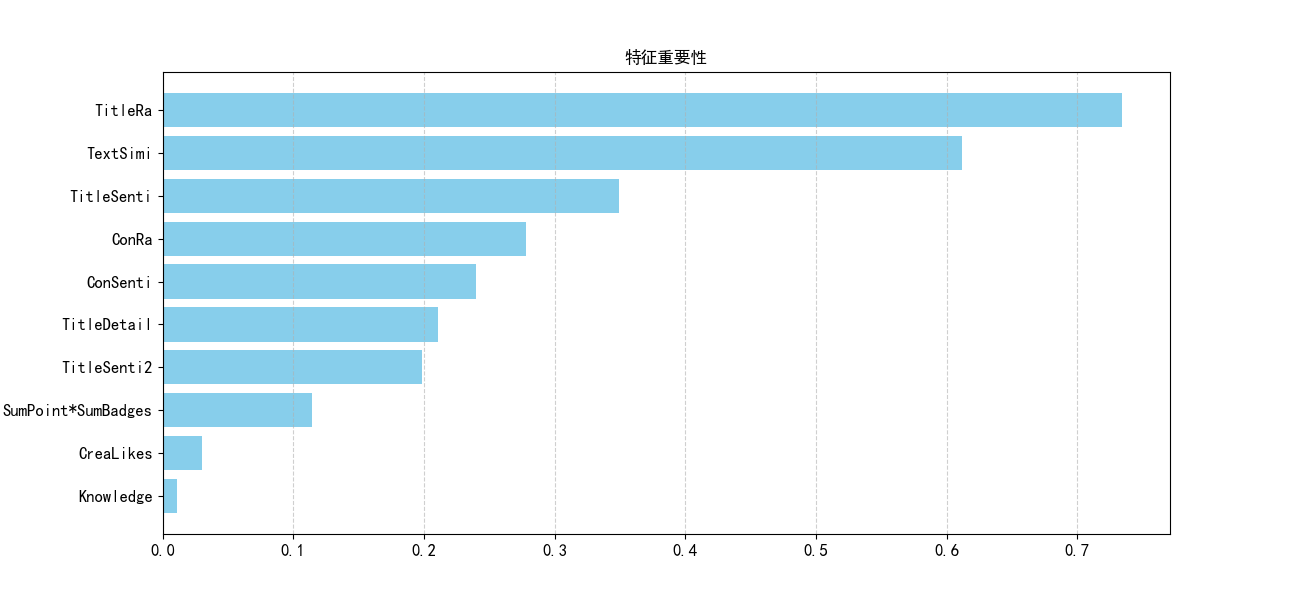


图 5.6 在线用户创新社区创意选择的特征重要性

Figure 5.6 Characteristics and Importance of Creative Choices in Online User Innovation Communities

在XGBoost算法中，通过对创意特征的重要性分析，可以深入了解影响创意评价的关键因素。首先，创意的标题可读性的重要程度是最高的，其重要性得分为0.73423，突显了标题在引起用户兴趣和理解创意方面的关键作用。其次，文本相似度、标题情感倾向以及内容可读性等因素也在创意选择中发挥了重要作用，分别占据不同程度的影响力。创意得分数、总得分乘以徽章数以及创意点赞数等社交互动指标对于创意选择同样具有显著意义。然而，技能标签数对创意评价的影响相对较小。

综合分析，在线用户创新社区的创意选择模型中，创意的大部分非结构化数据对创意选择都具有不同的重要程度，其中创意的标题可读性重要程度最高。

## 5.6 本章小结

本章建立了在线用户创新社区创意选择模型，并采用多种集成学习算法进行实验验证，筛选出XGBoost作为优化算法对创意进行选择，以提高创意选择的性能。通过SMOTE算法进行样本均衡，以及采用贝叶斯改进XGBoost算法的参数，得到的改进XGBoost算法在创意选择的过程中效果优于XGBoost算法。最后分析特征重要性，帮助企业了解在创意选择中哪些因素更具影响力，从而更合理地进行决策和资源分配。

# 第六章 结论与展望

本研究以在线用户创新社区创意为研究对象，构建创意价值的影响因素模型，分析创意价值影响因素，在此基础上构建了创意选择模型；研究本章是对整个研究的总结梳理，涵盖了本研究工作内容总结、结论、实践贡献及展望。首先，就本文主要工作进行回顾，依据前几章的研究得出研究结论；进一步，结合研究结论和实际情况，提出管理实践的建议，为社区管理者和社区用户带来一定的借鉴与启示；最后，总结研究局限性并提出未来的研究展望。

## 6.1 研究结论

企业建立在线用户创新社区主要是为了收集更多高质量的创意，以提高自身的创新能力。但是在线用户创新社区的创客具有不同的教育背景、文化水平，生成的创意层次不一，面对平台上创意信息过载的现状，如何及时有效地识别出高质量的创意是企业运行在线用户创新社区面临的一个巨大的挑战。这就需要企业深入的研究社区上创意和创客特征，以激励用户不断发布高质量创意，并有效地筛选出有价值的创意投入到企业自身产品研究和开发中。

因此本研究综合利用开放式创新理论和ELM模型构建了在线用户创新社区创意价值的影响因素模型，通过案例研究法、文本挖掘法、统计分析法等多种研究方法的综合使用，对在线用户创新社区的创意文本类型特征、创意规范类型特征、创客特征以及用户评论特征对创意价值的因果关系进行了全面系统地研究，并在得出结果的基础上，构建在线用户创新社区创意价值选择模型，通过改进XGBoost算法实现创意自动筛选效果。企业可以基于社区中创意和创客多方面的特征，建立创意评估系统，及时筛选出具有价值的创意，不但可以提高用户自我效能感，也可以提高企业创新效率。本文的研究结论主要归纳为以下方面：

（1）通过ELM模型，从中心路径假设和边缘路径假设探讨影响在线用户创新社区影响创意价值的因素。

在中心路径假设中，标题详细度对创意价值产生正向影响，较为详细的创意标题可以更清晰地传达创意的主要内容和特点，帮助企业或用户更好地理解创意，从而增加创意的价值。标题的可读性和内容的可读性对创意价值产生正向影响，即可读性高的标题更能吸引用户的注意力，并且可以更快地让企业和用户理解创意的主要内容，从而有助于提升创意的价值。可读性高的内容使企业和用户更轻松地理解创意，提高创意被接受和传播的可能性，因此对创意价值产生积极影响。此外，标题和内容的文本相似度也对创意价值产生正向影响，即标题和内容相似的创意可以保持一致的信息呈现，增强用户对创意的理解，从而有助于提升创意的价值。另外，标题和内容的情感倾向的平方值对创意价值产生负向影响，这表明带有主观情感的创意不会得到用户或企业的认可，而越客观的创意越具有创意价值，即创意标题和内容的情感倾向与创意价值呈倒U型关系。除此之外，研究结果显示，创意得分数对创意价值产生正向影响，即当一个创意得分较高时，通常意味着该创意在创新性和解决问题上表现出色，这使得它更有可能为企业或用户带来实际价值。创意得分越高，创意价值越大。创意点赞数也对创意价值有正向影响，更多的点赞意味着创意更受欢迎，从而提高创意的价值。

在边缘路径假设中，创客的徽章数正向影响创意价值，创客在在线用户创新社区中获得越多徽章表明其在社区中具有更高的影响力和创新能力，从而使其创意更受关注，提升创意的价值。创客的技能标签数正向影响创意价值，创客拥有更多的技能标签意味着其在多个领域拥有专业知识，这可以使其创意更具多样性和实用性，从而增加了创意的价值。不同领域的知识和技能可以为创意提供更丰富的背景和视角，有助于创意更好地解决现实问题或满足用户需求。创意获得的徽章数与创意总得分的交互作用会进一步强化创客的获得的徽章数与创意价值之间的正向关系。技能标签数与创意总得分的交互作用未显示对创意价值的显著影响，即创客的总得分并不正向调节创客的技能标签数对创意价值的影响。

（2）根据显著性影响因素结果构建创意选择模型，XGBoost集成学习算法对创意选择效果突出。通过在线用户创新社区创意价值的影响因素分析结果，得到多个关键特征，包括创意文本类型特征、创意规范类型特征和创客特征。在相同的数据集上，采用了Bagging和Boosting这两种常见的集成学习方法，包括RF、GBDT、AdaBoost、XGBoost以及LightGBM集成学习算法，利用Python内置包对上述五种算法进行了在线用户创新社区创意选择的建模工作。数据样本将被分为训练集和测试集两个部分输入到选择模型中，通过对比Precision、F1值、Recall和AUC的结果，研究发现，XGBoost算法在创意选择过程中表现更为出色。

（3）对XGBoost集成学习算法进行优化，创意选择效果优于XGBoost算法。为了解决样本不平衡的问题，使用了进行样本均衡处理。SMOTE通过生成合成的少数类别样本，有效提高了模型对于少数类别的学习能力，从而改善了整体的模型性能。其次，使用了贝叶斯优化方法对XGBoost的关键参数进行调参。在优化后的模型中，取得了显著的效果提升。精确度提高了约4.58%， F1值增加了约2.22%，召回率微幅提升约0.10%， AUC值有着显著的1.39%提高，进一步印证了模型在整体性能上的优越表现。通过综合运用SMOTE进行样本均衡和贝叶斯优化调参的手段，使XGBoost其在创新社区创意选择中取得了更为出色的性能，为实际应用中的创意选择任务提供了更有力的支持。

（4）在前人研究的基础上，本文致力于深化对非结构化创意文本的数据挖掘和情感分析。通过拓展在线用户创新社区数据采用范围，在研究中进一步完善了在线用户创新社区的研究框架。这一框架的完善不仅拓展了数据的来源，还提供了更全面的视角，以更好地理解和分析创新过程中的文本数据。另外，本文引入了文本相似度这一变量，通过对创意标题与创意内容的文本相似度的深入研究，其对创意价值之间的潜在关联，为在线用户创新社区上创意内容的优化提供更具体的指导。同时本文采用了多种集成学习算法，并将其应用到在线用户创新社区中，能够更有效地处理大规模的创意文本数据，提高创新社区的效率和可操作性。

## 6.2 管理启示

本文对Salesforce在线用户创新社区的研究结论可以为建立在线用户创新社区的企业在管理平台和用户，筛选社区高质量创意方面提供指导意见，特别是当平台创意信息过载时，同时企业又没有能力或者只有有限的人力和资源处理这些创意信息时。而且本文的研究结论也可以帮助企业选择出能够产生创意价值的创意。以下是根据本文的研究结论，提出的相应的在线用户创新社区的管理启示：

（1）本文可以帮助企业更好地完善创意提交方式，更好地管理创客提交的内容。首先之前的研究大多忽略了创意文本特征对创意价值的作用，而根据本文结论，包含创意文本特征的创意可能比不包含这些特征的创意更具有创意价值。因此企业在创意提交时可以提供创意文本模板，提醒创客应尽可能通俗地描述创意，使创意可读性更高，减少文本复杂度，在语言表达上不论是创意标题还是创意内容都需保持客观明确，尽量减少创意描述时用的情感词汇，这有利于企业和其他用户对创意的正确理解，减少用户理解创意的时间，提高创意价值和创意筛选效率。创意内容也应围绕创意标题描述，两者应具有相似度，这对提高创意价值更有帮助。再者企业可以为创客提供当前流行的创意主题信息，以便用户可以围绕流行的主题提交创意内容，以满足市场需求。企业还可以在用户提交创意时实时更新计算出的创意文本特征数据信息，以提醒用户完善其创意，从而获得更多的得分数，同时在平台上增加用户对创意可读性、创意内容流行度等特征的评价指标，以更直观的方式展现创意内容特征，减少创意筛选时的认知成本和搜寻努力。其次创客在提交创意时可以添加有说服力的图片以及标签信息，辅助描述创意内容，降低创意内容的模糊性。企业也要限制创意内容更新的上线，防止创客随意更新创意，降低创意可信度。

（２）本文还发现徽章数高以及具备先前经验的创客更具备产生创意价值的能力，因此企业识别出这类创客很重要。在开放式创新社区中，用户无法获得创意者的文化背景、教育经历等信息，但可以通过创客的徽章和总得分数据判断创客的能力，因此企业可以通过向创客总得分和徽章数的形式标志着具备产生创意价值的创客，以便用户和企业更容易地观察到这些创客提交的创意，同时也可以提高他们的参与度，促进用户创新。在创意选择模型的构建中发现，创客的特征对筛选高质量创意具有重要的影响，因此企业可以奖励那些发布了包含良好创意文本特征的创意的创客相应的得分和徽章，提高企业创意筛选能力和创新能力。

（３）现如今在线用户创新社区多以人工筛选创意为主，对企业人力和资源要求极高，如果需要等到创意达到一定数量才评定创意，则用户会因为等待时间过长而降低创意的积极性。因此企业需要基于创意特征和创客特征等多种变量，构建创意选择系统，以达到实时自动化筛选高质量创意的目的，有助于提高企业创新效率，更好地管理在线用户创新社区，同时促进用户持续参与。企业也可以利用创意选择机制将筛选出的创意优先展示给用户，减少用户信息搜索时间并降低不确定使用户和企业及时了解市场需求信息。

## 6.3 研究局限性及研究展望

本研究通过学习、总结以往在线用户创新社区相关研究，针对平台存在的信息过载、创意评定难度大等问题，探究在线用户创新社区创意价值的影响因素，并采用集成学习算法对创意进行筛选，有助于企业实现高效率地创意筛选，提高企业对平台和用户的管理能力。但是本研究也存在一定的不足之处，需要日后进一步改进：

（1）本研究仅采集Salesforce在线用户创新社区上的创意数据集，无法保证所研究的自变量对创意价值的影响机理在其他平台适用，特别是对于不同的语言和文化环境下的平台，这些平台中可能会有不同的特征对创意价值有影响。因此未来的研究可以选取其他的在线用户创新社区作为研究对象，进一步证明本研究结果的普适性。

（2）本研究的一个目的是探讨创意的文本特征对创意价值的作用，尽管已经对创意文本内容进行了深入的挖掘分析，但是仍然对创意内容分析不够全面，例如创意中是否包含相关链接、创意新颖性、创意内容的冗余度等没有考虑到模型中。

# 参考文献

[1] Henkel J, E V H. Welfare implications of user innovation [J]. The journal of technology transfer, 2005, 30: 73-87.

[2] 杨依依, 陈荣秋. 从封闭创新到开放创新——顾客角色、价值及管理对策 [J]. 科学学与科学技术管理, 2008, 29(03): 115-9.

[3] 陈钰芬, 陈劲. 用户参与创新: 国外相关理论文献综述 [J]. 科学学与科学技术管理, 2007, 28(2): 52-6.

[4] 高忠义, 王永贵. 用户创新及其管理研究现状与展望 [J]. 外国经济与管理, 2006, 28(4): 40-7.

[5] 姚山季, 王永贵. 顾客参与新产品开发对企业技术创新绩效的影响机制——基于B-B情境下的实证研究 [J]. 科学学与科学技术管理, 2011, 32(5): 34-41.

[6] E V H. Economics of product development by users: The impact of “sticky” local information [J]. Management science, 1998, 44(5): 629-44.

[7] 张裕伟. 用户创新的条件和范式研究 [D]. 浙江大学, 2008.

[8] Thomke S, E V H. Customers as innovators: a new way to create value[J]. Harvard business review [J]. Harvard business review, 2002, 80(4): 74-85.

[9] Hau Y S K Y G. Why would online gamers share their innovation-conducive knowledge in the online game user community? Integrating individual motivations and social capital perspectives [J]. Computers in Human Behavior, 2011, 27(2): 956-70.

[10] Sawhney M, Verona G, E P. Prandelli E. Collaborating to create: The Internet as a platform for customer engagement in product innovation [J]. Journal of interactive marketing, 2005, 19(4): 4-17.

[11] Hau Y S, M K. Extending lead user theory to users’ innovation-related knowledge sharing in the online user community: The mediating roles of social capital and perceived behavioral control [J]. International Journal of Information Management, 2016, 36(4): 520-30.

[12] Mahr D, A L. Virtual lead user communities: Drivers of knowledge creation for innovation [J]. Research policy, 2012, 41(1): 167-77.

[13] Fueller J, Muehlbacher H, Matzler K, et al. Consumer Empowerment Through Internet-Based Co-creation [J]. Journal of Management Information Systems, 2009, 26(3): 71-102.

[14] Rhyn M, Blohm I. A Machine Learning Approach for Classifying Textual Data in Crowdsourcing; proceedings of the 13th International Conference on Wirtschaftsinformatik, F, 2017 [C].

[15] Blohm I, Leimeister J M, H K. Crowdsourcing: How to Benefit from (Too) Many Great Ideas [J]. Mis Quarterly Executive, 2013, 12(4): 199-211.

[16] L B B. Crowdsourcing New Product Ideas over Time: An Analysis of the Dell IdeaStorm Community [J]. Management Science, 2013, 59(1): 226-44.

[17] Ma J, Lu Y, S G. User Innovation Evaluation: Empirical Evidence from An Online Game Community [J]. Decision Support Systems, 2019, 117: 113-23.

[18] Chesbrough H W. Open Innovation: The New Imperative for Creating and Profiting fromTechnology [M]. 2003.

[19] 陈劲, 陈钰芬. 开放式创新体系与企业技术创新资源配置 [J]. 科研管理, 2006, 27(3): 1-8.

[20] 梅鹏, 刘根节. 开放式创新范式下产学研协同的案例研究 [J]. 生态经济, 2013, (6): 131-4+45.

[21] 张振刚, 陈志明, 陈力恒. 企业开放式创新特征的归纳性分析—基于 28家创新型企业的问卷调查 [J]. 软科学, 2014, 12: 6-9.

[22] Muller A, Hutchins N, C P M. Applying Open Innovation Where Your Company Needs It Most [J]. Strategy & Leadership, 2012, 40(2): 35-42.

[23] Blohm I L J M, Krcmar H. Crowdsourcing: How to Benefit From (Too) Many Great Ideas [J]. Mis Quarterly Executive, 2013, 12(12): 199-211.

[24] Dahlander L, Frederiksen L, F R. Online Communities And Open Innovation [J]. Industry & Innovation, 2008, 15(2): 115-23.

[25] 秦敏, 李若男. 在线用户社区用户贡献行为形成机制研究:在线社会支持和自我决定理论视角 [J]. 管理评论, 2020, 32(09): 168-81.

[26] 张超. 支持开放式创新的企业研发平台功能及建设研究 [J]. 学习与实践, 2013, (12): 71-8.

[27] 金笑. 基于在线评论的顾客关键需求识别研究 [D]. 中国矿业大学.

[28] Alfarraj O, Alalwan A A, Obeidat Z M, et al. Examining the impact of influencers' credibility dimensions: attractiveness, trustworthiness and expertise on the purchase intention in the aesthetic dermatology industry [J]. International journal of commerce and management, 2021, (3): 31.

[29] Demir C. The Need for a Sustainable and Balanced Critical Approach in Book Reviews [J]. Discourse and Communication for Sustainable Education, 2019, 10(1): 20-36.

[30] Pornpitakpan C. The Persuasiveness of Source Credibility: A Critical Review of Five Decades' Evidence [J]. Journal of Applied Social Psychology, 2004, 34(2): 243-81.

[31] Fu X, Bang J, Kim M S. A Study on the Effects of CSR and Celebrity Model of Luxury Online Shopping Malls in China: Focusing on Elaboration Likelihood Model [J]. The Korea Academia-Industrial cooperation Society, 2021, (1).

[32] Kim H, Shawn Jang S. Restaurant-visit intention: do anthropomorphic cues, brand awareness and subjective social class interact? [J]. International journal of contemporary hospitality management, 2022, (6): 34.

[33] Shao Z, Zhang L, Pan Z, et al. Uncovering the dual influence processes for click-through intention in the mobile social platform: An elaboration likelihood model perspective [J]. Information & management, 2023.

[34] 左悦. 基于认知社会语言学的汉语网络热词研究 [D]. 吉林大学, 2023.

[35] Foell P, Thiesse F. Exploring Information Systems Curricula: A Text Mining Approach [J]. Wirtschaftsinformatik, 2021, (6): 63.

[36] 张宁, 聂嘉豪, 庞智亮. 基于在线评论的客户敏捷性与产品绩效的关系 [J]. 管理科学, 2022, 35(2): 47-59.

[37] 钱明辉, 徐志轩, 王玉玺. 基于网络招聘信息文本挖掘的企业竞争力识别研究 [J]. 管理评论, 2022, 34(07): 150-6.

[38] 朱青, 卫柯臻, 丁兰琳, et al. 基于文本挖掘和自动分类的法院裁判决策支持系统设计 [J]. 中国管理科学, 2018, 26(01): 170-8.

[39] Hienerth C, E V H. Berg Jensen M. User Community Vs. Producer Innovation Development Efficiency: A First Empirical Study [J]. Research Policy, 2014, 43(1): 190-201.

[40] J. Y H, A K. Investigating The Antecedents Of Organizational Task Crowdsourcing [J]. Information & Management, 2015, 52(1): 98-110.

[41] Dong J Q, W W. Business Value of Social Media Technologies: Evidence From Online User Innovation Communities [J]. The Journal of Strategic Information Systems, 2015, 24(2): 113-27.

[42] 李奕莹, 戚桂杰. 创新价值链视角下企业在线用户创新社区管理的系统动力学研究 [J]. 商业经济与管理, 2017, (06): 60-70.

[43] 吴冰, 卢彦君. 开放式创新社区对企业创新绩效的影响 [J]. 知识管理论坛, 2019, 4(04): 205-17.

[44] Von Briel F, J R. Lessons From A Failed Implementation Of An Online Open Innovation Community In An Innovative Organization [J]. Mis Quarterly Executive, 2017, 16(1): 35-46.

[45] Chang D, C C. Product Concept Evaluation And Selection Using Data Mining And Domain Ontology In A Crowdsourcing Environment [J]. Advanced Engineering Informatics, 2015, 29(4): 759-74.

[46] 程巧莲, 尹立国, 孙永川. 企业主导的创新社区构建:开放式创新的视角 [J]. 科研管理, 2017: 487-93.

[47] 董洁林, 陈娟. 无缝开放式创新:基于小米案例探讨互联网生态中的产品创新模式 [J]. 科研管理, 2014, 35(12): 76-84.

[48] 黄璐. 开放式创新社区创意采纳预测的研究 [D]. 华南理工大学, 2021.

[49] Chen L, Marsden J. R, Zhang Z J. Theory and Analysis of Company-Sponsored Value Co-Creation [J]. Journal of Management Information Systems, 2012, 29(2): 141-72.

[50] Lee H M, W V D. Creative Participation: Collective Sentiment In Online Co-Creation Communities [J]. Information & Management, 2015, 52(8): 951 -64.

[51] J V B F V R. Lessons from a Failed Implementation of an Online Open Innovation Community in an Innovative Organization [J]. Mis Quarterly Executive, 2017, 16(1): 35-46.

[52] Pirkkalainen H, Pawlowski J M, M B. Engaging In Knowledge Exchange: The Instrumental Psychological Ownership In Open Innovation Communities [J]. International Journal Of Information Management, 2018, 38(1): 277-87.

[53] 李悦. 在线创新社区中创新激励对用户创新意愿的影响研究 [D]. 东北财经大学, 2020.

[54] 赵婧. 小米社区用户贡献性参与行为的影响因素研究 [D]. 大连理工大学, 2022.

[55] 刘伟, 张波椋. 企业自建在线创新社区中反馈对用户持续创新贡献的影响研究——新产品开发阶段的调节效应 [J]. 系统工程理论与实践, 2022, 42(10): 2769-81.

[56] 詹湘东. 基于用户创新社区的开放式创新研究 [J]. 中国科技论坛, 2013, (08): 34-9.

[57] Huang Y, Singh P V, K S. Crowdsourcing New Product Ideas Under Consumer Leaming [J]. Management Science, 2014, 60(9): 2138-59.

[58] Brem A, V B. The Search For Innovative Partners In Co-Creation: Identifying Lead Users In Social Media Through Netnography And Crowdsourcing [J]. Journal of Engineering And Technology Management, 2015, 37: 40-51.

[59] 唐洪婷, 李志宏, 秦睿. 基于超网络的大众协同创新年社区用户知识模型研究 [J]. 管理学报, 2017, 14(6): 859-67.

[60] 王莉, 李沁芳, 马云龙. 基于改进网络志方法的开放式创新社区中领先用户识别研究 [J]. 科研管理, 2019, 40(10): 259-67.

[61] 单晓红, 王春稳, 刘晓燕, et al. 开放式创新社区领先用户识别——知识基础观视角 [J]. 数据分析与知识发现, 2021, 5(06): 85-96.

[62] Hossain M, Islam K M Z. Generating Ideas on Online Platforms: A Case Study of "My Starbucks Idea" [J]. Arab Economic & Business Journal, 2015, 10(2): 102-11.

[63] Mokter H, Zahidul I K M. Ideation Through Online Open Innovation Platform: Dell Ideastorm [J]. SSRN Electronic Journal, 2015.

[64] Rauter R, Perl-Vorbach E, Baumgartner R J. Is open innovation supporting sustainable innovation? Findings based on a systematic, explorative analysis of existing literature [J]. International Journal of Innovation & Sustainable Development, 2017, 11(2-3): 249.

[65] Li M, Kankanhalli A, Kim S H. Which ideas are more likely to be implemented in online user innovation communities? An empirical analysis [J]. Decision Support Systems, 2016, 84(Apr.): 28-40.

[66] Schemmann B, Herrmann A M, Chappin M M H, et al. Crowdsourcing ideas: Involving ordinary users in the ideation phase of new product development [J]. Research Policy, 2016, 45(6): 1145-54.

[67] Wang N, Tiberius V, Chen X, et al. Idea selection and adoption by users – a process model in an online innovation community [J]. Technology Analysis and Strategic Management, 2020, (4): 1-16.

[68] Wang W, Liu Y. Does University-industry innovation community affect firms' inventions? The mediating role of technology transfer [J]. The Journal of Technology Transfer, 2021, 47: 906 - 35.

[69] 郭琴. 创新源分散化条件下开放式创新社区创意价值识别研究 [D]. 广西大学, 2017.

[70] 王婷婷. 创新价值链视角下企业开放式创新平台创意管理研究 [D]. 山东大学, 2018.

[71] 李悦. 在线创新社区中创新激励对用户创新意愿的影响研究 [J].

[72] 高贝伦. 开放式创新社区创意采纳的标题影响因素研究 [D]. 广西大学, 2019.

[73] Klein M, Bicharra Garcia A C. High-speed idea filtering with the bag of lemons [J]. Decision Support Systems, 2015, 78(OCT.): 39-50.

[74] 祁运丽. 在线用户创新社区中创意采纳的影响因素研究 [D]. 北京工商大学.

[75] 王松, 杨洋, 刘新民. 基于图注意力网络的开放式创新社区用户创意潜在价值发现研究 [J]. 数据分析与知识发现, 2021, 5(11): 13.

[76] 马玉. 基于在线评论数据挖掘的产品创新机会识别方法研究 [D]. 吉林大学, 2023.

[77] 薛水晶, 郭伟, 连文涛, et al. 基于机器学习的开放创新社区创意识别方法研究 [J]. 包装工程, 2023, 44(08): 72-80.

[78] 陶晓波, 徐鹏宇, 樊潮, et al. 创新社区中新产品开发人员信息采纳行为的影响机理研究 [J]. 管理评论, 2020, 32(10): 12.

[79] Park Y E. The Effects of Message Framing and Source Credibility in Cause-Related Marketing [J]. Asian Journal of Public Relations, 2021, 4(1).

[80] 刘伟, 徐鹏涛. O2O电商平台在线点评有用性影响因素的识别研究——以餐饮行业O2O模式为例 [J]. 中国管理科学, 2016, 24(05): 168-76.

[81] Wu, Bing. Influence of MOOC learners discussion forum social interactions on online reviews of MOOC [J]. Education and Information Technologies, 2021, 26(3): 3483-96.

[82] 姚晓彤. 高校学生阅读行为类型及影响因素研究 [J]. 图书馆杂志, 2021, 40(03): 49-55.

[83] Kara A, H P. Assessment of the readability of online texts related to specific learning disorder [J]. Medicine Science | International Medical Journal, 2020, 9(1): 114.

[84] Xiaobin Chen, Meurers D. Linking text readability and learner proficiency using linguistic complexity feature vector distance [J]. Computer Assisted Language Learning, 2019, 32(4): 418-47.

[85] Siering M, Muntermann J, B R. Explaining and Predicting Online Review Helpfulness: The Role of Content and Reviewer-Related Signals [J]. Decision Support Systems, 2018, 108: 1-12.

[86] Zhang, Wei, Watts, et al. Capitalizing on Content: Information Adoption in Two Online communities [J]. Journal of the Association for Information Systems, 2008, 9(2): 73-94.

[87] 辛冲, 李明洋. 创新生态系统中知识基础关系特征与价值共创 [J]. 管理科学, 2022, 35(02): 16-31.

[88] Klára B. Using a parallel corpus to adapt the Flesch Reading Ease formula to Czech [J]. Journal of Linguistics/Jazykovedný casopis, 2021, 72(2): 477-87.