

学士学位论文

电商平台中互补者的产品同质特征与销量趋势预测研究

学 号： 20191000960

姓 名： 徐嘉艺

学 科 专 业： 信息管理与信息系统

指 导 教 师： 朱镇 教授

培 养 单 位： 经济管理学院

二〇二三年五月

中国地质大学（武汉）学士学位论文原创性声明

本人郑重声明：本人所呈交的学士学位论文《电商平台中互补者的产品同质特征与销量趋势预测研究》，是本人在指导老师的指导下，在中国地质大学（武汉）攻读学士学位期间独立进行研究工作所取得的成果。论文中除已注明部分外不包含他人已发表或撰写过的研究成果，对论文的完成提供过帮助的有关人员已在文中说明并致以谢意。

本人所呈交的学士学位论文没有违反学术道德和学术规范，没有侵权行为，并愿意承担由此而产生的法律责任和法律后果。

学位论文作者签名：_____

日 期： 2023 年 5 月 21 日

摘要

在互联网高速发展、信息高度透明的背景下，电商平台中的机会与挑战并存，企业面临着更复杂的竞争关系。互补者在入驻平台的过程中，既希望“求同”以顺利进入市场，又希望“存异”打造差异化竞争，增强自身竞争优势。虽然关于互补者在平台中绩效预测的研究并不少见，但鲜有研究探究产品特征对企业绩效的作用；因此，本研究基于携程平台中的大规模产品数据，关注了平台所有者在平台中的强大影响力，从互补者与平台产品的同质特征视角，探究了互补者的产品特征对其销量趋势预测的作用与价值。

本研究以携程平台中的大尺度旅游产品数据为对象，首先将产品分解为十个维度，使用命名实体识别技术将标题文本进行实体抽取及整合；其次，本文构建了供应商-产品图网络结构，使用图节点相似度算法衡量了每个时期内互补者与携程平台在每个产品维度下的同质化水平；接着，本文从产品特征、互补者与平台的产品同质化水平、互补者自身特征和市场竞争情况四个角度整理了 42 个相关变量，基于 XGBoost 训练了互补者的销量趋势预测模型并对模型进行了稳健性检验，最后使用 SHAP 对销量趋势预测模型进行了可解释性分析。

研究结论表明，与平台产品的同质化水平在互补者未来销量趋势预测中起到了重要作用。其中，酒店服务、供应商服务和景点这三个维度的指标贡献价值超过了 70% 的变量。在相关性分析中本研究还发现（1）在不同细分市场中，产品每个维度的同质化水平对预测模型的贡献程度是不同的，主要和产品各维度内容在不同细分市场下的丰富度不同有关；（2）产品规模不同的互补者在销量趋势预测中对产品同质化水平指标的依赖性是不同的，产品规模较大的互补者更依赖传统的销量趋势预测指标，而产品规模较小的互补者除了依赖传统指标之外，也在很大程度上依赖于产品的同质化水平指标。

本研究的创新点为：（1）关注了产品自身特征对企业绩效的预测作用，提出了旅游电商产品的特征维度定义方法，基于产品同质化水平的视角探究了产品本身特征对企业绩效的预测作用；（2）提出了一种基于图网络与图计算的产品同质化水平的衡量方法，挖掘了产品文本更细粒度的信息，更全面地从宏观和微观两个角度来衡量产品的同质化水平。

关键词：平台互补者；旅游产品；产品同质化程度；图网络；可解释机器学习

Abstract

In the context of the rapid development of the Internet and highly transparent information, opportunities and challenges coexist in the e-commerce platform, and enterprises are faced with more complex competitive relations. In the process of entering the platform, the complementator not only hopes to "seek common ground" to enter the market smoothly, but also hopes to "set aside differences" to create differentiated competition and enhance their own competitive advantages. Although the research on the performance prediction of the complementator in the platform is not rare, few studies have explored the effect of product characteristics on enterprise performance. Therefore, based on the large-scale product data in the Ctrip platform, this study focused on the strong influence of platform owners in the platform, and explored the role and value of the product characteristics of the complementator on its sales trend prediction from the perspective of the homogeneity of the complementator and the platform products.

Large-scale tourism product data from Ctrip platform was taken as the object in this study. Firstly, the product was decomposed into ten dimensions, and the named entity recognition technology was used to extract and integrate the title text. Secondly, this paper constructs the vendor-product graph network structure, and uses the graph node similarity algorithm to measure the homogeneity level of the complementator and Ctrip platform in each product dimension in each period. Then, this paper sorts out 42 related variables from four perspectives: macro factors, product homogeneity level of the complementator and platform, the complementator's own characteristics and market competition. Based on XGBoost, the sales trend prediction model of the complementator is trained and the model is tested for robustness. Finally, SHAP is used to analyze the interpretability of the sales trend prediction model.

The results show that the homogenization level of products with the platform plays an important role in predicting the future sales trend of the complementator. Among them, the three dimensions of hotel service, supplier service and scenic spots contribute more than 70% of the variables. In the correlation analysis, it is also found that (1) in different market segments, the homogenization level of each dimension of the product makes different contribution to the prediction model, which is mainly related to the different richness of each dimension of the product in different market segments. (2) Complementors with different product sizes have different dependence on product

homogeneity level index in sales trend prediction. Complementors with larger product scale are more dependent on traditional sales trend prediction index, while complementors with smaller product scale are also largely dependent on product homogeneity level index in addition to traditional index.

The innovation points of this study are as follows: (1) Focusing on the predictive effect of product characteristics on enterprise performance, it proposes the definition method of characteristics dimension of tourism e-commerce products, and explores the predictive effect of product characteristics on enterprise performance from the perspective of product homogeneity; (2) A measure method of product homogeneity level based on graph network and graph calculation is proposed to mine finer grained information of product text and measure product homogeneity level from macro and micro perspectives in a more comprehensive way.

Key words: platform complementator; Tourism products; Degree of product homogeneity; Graph network; Explainable machine learning

目 录

第一章 绪论	1
1.1 研究背景与研究问题	1
1.2 研究目的与研究意义	3
1.2.1 研究目的	3
1.2.2 研究意义	3
1.3 研究创新点	3
1.4 论文结构	4
1.5 技术路线	5
第二章 文献综述与理论基础	6
2.1 平台与互补者	6
2.1.1 平台生态系统	6
2.1.2 平台与互补者的竞争	7
2.2 产品同质化水平及测量	8
2.2.1 产品同质化水平的概念	8
2.2.2 产品同质化水平的测量	8
第三章 研究方法	10
3.1 命名实体识别	10
3.1.1 命名实体识别的定义	10
3.1.2 命名实体识别的实现技术	11
3.2 图结构与节点相似度计算	12
3.2.1 图结构定义与构建	12
3.2.2 节点相似度算法	13
3.3 XGBoost 模型	14
3.3.1 XGBoost 简介	14
3.3.2 XGBoost 模型原理	14
3.4 机器学习模型的可解释性方法	16
3.4.1 可解释性理论	16
3.4.2 基于 SHAP 的可解释分析方法	17
第四章 数据与变量测量	18
4.1 数据	18
4.2 产品特征提取和相似度计算	19
4.2.1 旅游产品特征维度定义	19

4.2.2 产品特征抽取	22
4.2.3 互补者与平台间的产品同质化水平计算	22
4.3 变量定义	24
4.3.1 产品特征	25
4.3.2 互补者与平台间的产品同质化水平	26
4.3.3 互补者特征	27
4.3.4 市场竞争情况	28
第五章 销量趋势预测模型	30
5.1 数据集预处理	30
5.2 模型选择	30
5.3 模型训练及参数调整	31
5.4 模型结果评估	32
5.5 稳健性检验	33
第六章 基于 SHAP 的模型解释性分析	34
6.1 特征重要性分析	34
6.2 目的地城市相似度与模型预测效果	36
6.3 特征相关性分析	38
6.3.1 细分市场类别与产品同质化水平	38
6.3.2 互补者规模与产品相同质化水平	39
第七章 结论与展望	41
7.1 主要研究结论	41
7.2 管理建议	41
7.3 不足与展望	42
致谢	43
参考文献	44
附录 A: 命名实体识别过程的技术细节	48
附录 B: 基于 Neo4j 的供应商-产品网络可视化结果	50
附录 C: 变量统计汇总数据	51

第一章 绪论

平台生态的出现不断吸引着大量中小微型企业的加入，这些通过为平台提供互补品而为平台创造价值的企业被称为互补者。在平台系统的信息高度透明的背景下，机会与挑战共存，在广阔的平台带来了更多市场机会的同时，竞争对手的学习和模仿成本也更低。在这种情况下，互补者在市场中的生存则面临着矛盾之处：一方面，互补者希望“求同”以更顺利地进入市场；另一方面，互补者又希望“存异”来差异化竞争。在这一过程里，平台对互补者的影响是巨大的，因为平台具有大量的资源、价格等优势。如今，关于互补者在平台中的生存与绩效的研究已并不少见，如何从互补者的产品本身特征出发，关注平台所有者在平台生态中的强大影响力，研究互补者在平台中的绩效表现将有助于互补者更好地在平台竞争中求得先机。

1.1 研究背景与研究问题

数字经济的出现推动了平台经济的发展，从而引发越来越多的中小微型企业纷纷将自身业务搬上互联网的舞台。中国的电商规模庞大，2022年市场规模已达到了31.4万亿元，同比2021年提高了7.86%¹。中国的电商总体增势强，规模已经远超过了传统市场。在庞大的市场规模下，平台生态系统中亦存在复杂而激烈的多方竞争关系。基于平台的电子商务市场通常由多个不同的参与者组成，其中一个典型的特征是平台所有者和互补者之间的标准化和多边关系^[1]。由于平台生态系统的涌现创造了大量的商业机会，在互联网快速发展和信息高度透明共享的背景下，这种大量的商业机会通常伴随着激烈的竞争^[2]。竞争通常来自多方，例如互补者之间的竞争以及平台和互补者之间的竞争^[3]。平台与互补者之间竞争关系的影响并不是绝对的，尽管平台通常在互补者面前展现出强大的资源、价格优势。大部分情况下互补者的体量规模相对平台来说较小，一般依靠平台提供的资源而实现生存^[4]。但如果平台依靠其强大的资源优势不顾后果地与互补者开展竞争的结果是可怕的，这将很可能导致平台失去活性——因为平台的成功标准并不

¹ 数据来自《2022年度中国产业电商市场数据报告》

是一家独大，而是取决于吸引消费者和互补者加入的能力。在这种情况下，平台的运作者将要考虑更多因素来维护自身的运行。

在旅游市场中，供应商往往基于自身旅游资源来决定产品的设计与投放，这些旅游资源包括供应商所持有的目的地城市资源、景点资源以及酒店资源等等。在旅游电商平台中，平台所有者一般手持大量旅游资源，这些旅游资源可以允许平台进行大规模的产品投放以及快速的产品迭代与更新，从而拥有更强的产品创新和产品模仿的能力。而对于互补者来说，其旅游资源相对于平台来说较为匮乏：通常聚焦于单一的旅游细分市场或特定的业务。那么在这种情况下，平台是否已经主导了竞争地位呢？答案很可能是，因为平台强势的竞争地位将会降低其他互补者的活性，进而可能导致平台走下坡路。因此，适度的竞争环境对平台和互补者来说都是一个有利因素。

供应商通常通过产品间的模仿、创新等行为来增强自身的竞争优势，对于这种模仿、创新行为的判定是理解竞争关系的关键。例如，Wang 等使用了移动应用的发布时间来衡量手机应用的模仿者与创新者^[5]。其探究了竞争环境下模仿行为对于原创产品的影响。鉴于旅游产品的特殊性，本研究很难像其他实体产品一样明确定义创新者的角色，从而判断出某个产品或者某个供应商的绝对创新性及与其他产品或供应商是否存在模仿关系。因此，两个供应商所投放产品的同质化水平的提升也被我们认为是增加其竞争激烈程度的一种。例如，Huang 在针对淘宝网产品的研究中发现，产品描述越相似的可替代性产品会加剧市场内的竞争。因此这会对焦点产品产生消极影响^[6]。在以往的研究中，大多数研究通常基于文本数据使用单纯的文本相似度计算来衡量产品同质化水平，例如 Huang 在研究中使用了 TF-IDF 和余弦相似度来衡量产品文本之间的相似度，很少有研究从产品特征的细粒度层面，将产品看作现实世界中的实体以作进一步研究。而旅游产品作为一种特殊的体验类产品，我们需要适当的方法对其进行结构化表征以衡量供应商之间的产品同质化水平。Smith 在 1994 年首次将旅游产品分解为五个元素^[7]。分别为：物理实体、服务、好客程度、选择自由度和游客参与程度；Xu 则在 2010 年基于新视角的基础上对其进行了改进。其将后四个元素视为平等关系，而非 Smith 提到的包含关系^[8]。本文在此基础上，对于旅游电商平台上的旅游产品进行结构化分类以期达到对旅游产品细粒度的结构化测量与分析效果。

因此，本文研究的具体问题是：在旅游电商平台中，对于互补者来说，与平台所有者之间的产品同质化水平是否可以成为预测其未来销量趋势的重要指标？如果可以，那么这种指标对预测模型的贡献大小与哪些因素有关？

1.2 研究目的与研究意义

1.2.1 研究目的

基于上述的研究背景介绍，本文的研究目的是：（1）探究一种将产品进行实体化表达以反映微观产品特征的方法，并在此基础上提出一种考虑到市场全局的宏观视角与产品本身的微观视角下的产品同质化水平测量方法；（2）研究与平台的产品同质化水平是否可以作为互补者未来销量趋势预测的关键指标，并对比其与其他传统预测指标（价格、口碑、销量等）对预测模型贡献价值的大小。

1.2.2 研究意义

本研究的研究意义如下：

（1）探究了电商平台中互补者的产品特征对其销量趋势预测的重要作用。在传统销量趋势预测指标（价格、口碑、历史销量等）之外，关注了产品特征在销量趋势预测中的重要贡献，揭示了互补者自身产品策略与绩效表现的关系，为互补者在复杂的平台竞争中取得优势提供了新的策略角度。对互补者来说，互补者在进行市场定位或制定产品策略时，应全面考虑到平台的多方复杂竞争关系，尤其是小规模互补者，要避免与平台所有者的产品同质化水平过高，以失去竞争优势。对平台来说，虽然其在平台生态中占据较大的优势，但也应顾及平台整体生态，防止过度侵占互补者市场导致互补者活性降低。

（2）提出了基于供应商-产品网络的产品同质化特征计算方法，使用了命名实体识别技术将产品进行实体化表达，挖掘了产品的微观特征，丰富了以往有关产品同质化特征或相似度的相关研究。在进行产品同质化水平测量时，可以考虑到市场全局的宏观角度来提升产品同质化水平测量的准确性。有助于平台对产品的规范化治理，也为企业的产品定位与竞争提供了数据策略层面的参考。

1.3 研究创新点

本研究的主要创新点如下：

一、本研究关注了产品自身特征对企业绩效的预测作用，提高了企业预测未来销量趋势的准确性：本研究在多数研究中关注的企业绩效预测的传统特征（销量、价格、口碑）之外，提出了旅游电商产品的特征维度定义方法，使用了命名

实体识别技术将产品进行实体化表达，挖掘了产品本身所蕴含的丰富信息，研究了产品同质特征对企业绩效预测的作用，提升了企业对未来销量趋势预测的准确性。

二、本研究提出了一种基于图网络计算的产品同质化程度的衡量方法，提高了产品同质化水平测量的精确度：以往的研究大多数从整体角度，以 TF-IDF 或余弦相似度等方法来衡量出两个产品文本之间的相似度。这种方法很容易忽略整体市场的情况，只将相似度计算局限于两个参与计算的产品之间。而本研究首先使用 BERT-LSTM-CRF 的自然语言处理模型对旅游产品的大规模文本进行抽取，然后将其构建为图网络结构来进行相似度计算。由于图网络中富含着丰富的点、边、权重等信息，可以更好地从市场全局的宏观视角来衡量产品的同质化水平，提升了产品同质化水平测量的精确度。

1.4 论文结构

本文的主要研究对象是携程平台中的平台互补者，希望从与平台所有者的产品同质化水平的视角下，研究互补者的产品特征对互补者销量趋势预测的作用。本文首先定义了旅游电商产品的产品特征维度，并使用自然语言处理技术进行命名实体识别，然后构建供应商-产品网络来计算互补者与平台产品的同质化程度。最后，基于 XGBoost 模型建立了互补者销量预测模型，并使用 SHAP 进行模型可解释性分析。

本文一共分为以下七个章节和附录部分：

第一章，绪论。本章介绍了本研究的研究问题提出背景与具体的研究问题描述，阐述了研究目的与意义，最后总结了整体研究设计、研究路线和技术路线。

第二章，文献综述与理论基础。本章主要介绍了研究建立所依靠的先前研究与理论。介绍了平台与互补者的过往研究方向和进展，总结了历史关于产品同质化水平相关的研究内容以及计算方法。

第三章，研究方法。本章阐述了本研究中使用到的各类技术方法，包括命名实体识别技术、图结构与节点相似度计算、XGBoost 模型原理以及机器学习模型的可解释性方法。

第四章，数据与变量测量。本章首先介绍了产品特征提取与相似度计算的具体过程，然后总结了用于构建互补者销量预测模型的四类变量，主要包括在第四章得出的控制变量——产品同质化水平变量——以及额外纳入考虑的三类变量，分别是产品特征变量、互补者特征变量以及市场竞争情况变量。

第五章，销量趋势预测模型。本章描述了互补者销量趋势预测模型建立的具体过程，包括模型的选择与对比、数据集处理、模型训练及参数调整以及模型结果评估与稳健性检验。

第六章，基于 SHAP 的模型解释分析。本章基于 SHAP 对第六章训练的销量趋势预测模型进行了特征可解释性分析，主要包括特征重要性分析和特征相关性分析；

第七章，结论与展望。本章概述了研究的主要结论与内容，阐述了本研究的具体管理意义，最好提出了研究的不足与展望。

附录。附录 A 记录了本文第四章中进行产品特征提取所运用的命名实体识别技术的具体细节，包括文本标注过程、深度学习模型参数设置以及模型结果评估等；附录 B 展示了本研究基于 Neo4j 数据库建立的供应商-产品网络可视化结果；附录 C 记录了用于训练 XGBoost 模型的变量的统计汇总数据。

1.5 技术路线

根据研究内容，本文的技术路线图如图 1.1 所示：

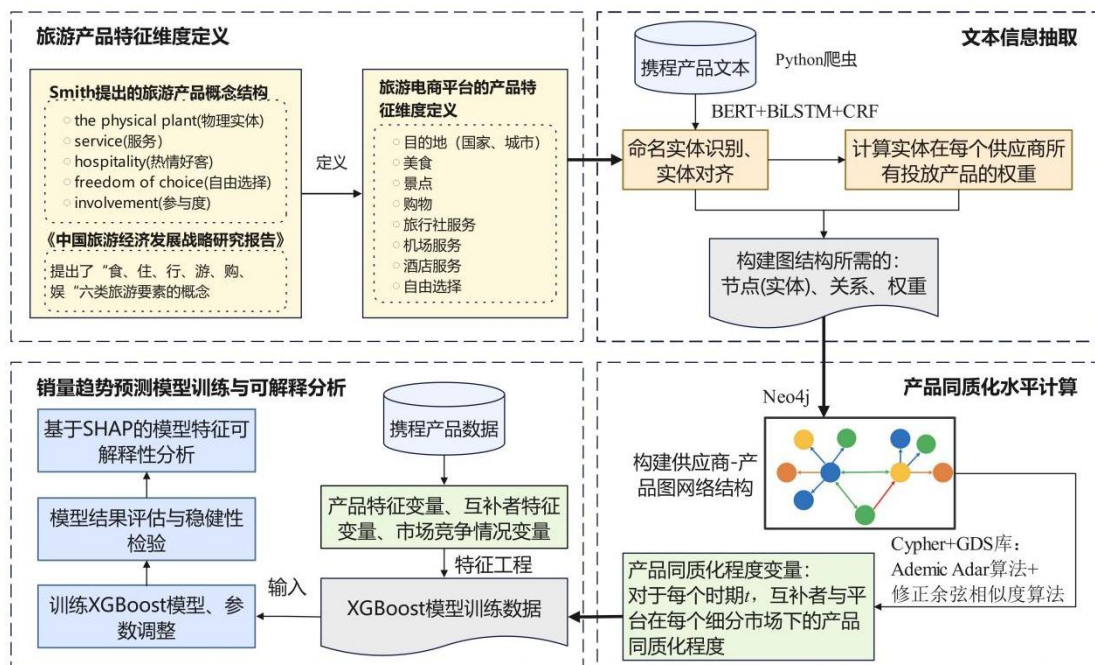


图 1.1 技术路线图

第二章 文献综述与理论基础

本章阐述了平台与互补者以及产品同质化水平相关的文献综述和理论研究，旨在为本研究提供研究支持与理论基础。本章主要包括以下内容：2.1 节介绍了平台与互补者相关的概念，包括平台生态系统的概念介绍与生态机制解释以及平台与互补者的竞争关系研究及进展梳理；2.2 节介绍了产品同质化水平相关的概念及计算，包括过去研究如何利用产品同质化水平进行商业研究以及有关产品同质化水平的计算方法整合。总体来说：（1）目前历史研究中关于互补者与平台之间竞争关系的研究多聚焦于平台自身策略、互补者战略及行为层面，很少有研究细粒度聚焦产品本身，并以此反映供应商在电商平台中的产品投放策略；（2）多数关于产品同质化水平的计算方法多为基于文本的 TF-IDF 或余弦相似度计算，而将产品看作由一系列特征属性构成的实体，然后基于图网络来研究产品同质化程度的研究仍有空缺。

2.1 平台与互补者

2.1.1 平台生态系统

有关平台生态系统的概念最早来自于商业生态系统理论，商业生态系统的概念最早于 1993 年由 Moore 提出^[9]。Moore 将商业生态系统看作由多方组成的经济联合体，多方则包括诸如生产者、政府、行业协会及其他的利益相关者等，该概念的提出灵感源于生态学中“关键物种”的概念。Moore 认为平台所有者是商业生态系统中的关键型企业，关于平台的研究与商业系统的研究越来越紧密结合，平台型商业生态系统（Platform-based business ecosystem）一词也逐渐推广开来。陈威如等人也对平台型生态系统进行了定义^[10]。其研究认为，柔性及动态的价值创造是平台生态系统的基本特征。在平台核心价值驱动下，平台生态系统即为构建出的多边、合作、共赢机制。Iansiti 指出，平台一方面吸引消费者的加入，另一方面也吸引互补者的入驻，而平台的成功则取决于吸引他们的能力。在这种情况下，平台的运作者将要考虑更多因素来维护自身的运行。平台需要维护整个生态系统的健康因为他们的生存依赖于此^[11]。

平台生态系统的仍属于商业生态系统的一种。其既包括平台的特征，也包括平台的功能和属性。基于平台的电子商务市场通常由多个不同的参与者组成，其

典型的特征则是平台所有者和互补者之间的标准化和多边关系。除了多边性之外，平台还具有网络效应，即平台促进不同群体间的交易或交互，既需要吸引平台使用者（例如消费者）的加入，也需要吸引互补企业的加入。平台所有者在满足供给和需求端的需求同时，也要通过协调组织、共享资源和关系治理等方式来平衡平台各方利益。

目前关于平台生态系统相关的研究多聚焦于平台所有者。例如平台治理^[12]：一些研究关注于平台所有者和互补者之间的决策分配^[13]、平台所有者的正式和非正式控制机制以及平台所有者的跨界行为和包络战略等^[14]。另外，也有研究聚焦于平台进入问题。例如，Gawer 研究了英特尔公司对互补者市场的行为^[15]。其研究认为当英特尔会选择加入互补者市场来激励他们的创新。Javier 建议互补者积极利用平台的各种资源和自身的独特优势来增强竞争力^[16]。Li 等人的研究也揭示了平台对互补者的激励关系^[17]。其研究表明，平台与互补者竞争关系的出现还会引发注意力溢出效应，从而吸引消费者的更多关注并在一定程度上激发互补者的创新能力。

2.1.2 平台与互补者的竞争

基于平台的电子商务市场通常由多个不同的参与者组成，其一个典型的特征则是平台所有者和互补者之间的标准化和多边关系。互补者进入平台生态系统的目的是销售其所提供的产品或服务给平台的终端用户^[18]。平台生态系统的涌现伴随着大量的商业机会，因此互补者在大量商业机会面前，也面临着激烈的竞争。竞争通常来自多方，例如互补者之间的竞争以及平台和互补者之间的竞争。平台与互补者之间竞争关系的影响不是绝对的，尽管平台通常在互补者面前展现出强大的资源、价格优势，而大部分情况下互补者的体量规模相对平台来说较小，一般依靠平台提供的资源而实现生存^[5]，但如果平台依靠其强大的资源优势不顾后果地与互补者开展竞争的结果是可怕的，这将很可能导致平台失去活性——因为平台的成功标准并不是一家独大，而是取决于吸引消费者和互补者加入的能力。在这种情况下，平台的运作者将要考虑更多因素来维护自身的运行。

互补者对平台生态系统的价值创造具有重要的意义^[19]。历史研究中，已经有许多研究聚焦于平台与互补者的竞争关系，许多研究致力于探索平台促进互补者的活性，例如一些研究表明平台与互补者间的适度竞争是有益的。例如 Gawer 发现，当英特尔公司对互补者市场不满意时就会选择加入市场来激励他们的创新^[14]。Javier 建议互补者积极利用平台的各种资源和自身的独特优势来增强竞争力^[15]。Li 等人表明，平台与互补者竞争关系的出现还会引发注意力溢出效应^[16]。这种注

意力溢出效应可以吸引消费者的更多关注并在一定程度上激发互补者的创新能力。然而，大多数有关平台与互补者竞争的研究更多聚焦在企业的自身行为层面，例如有关平台进入或互补者自身创新或产品营销等方面。很少有研究关注产品特征的微观层面，例如产品的同质化。未来，借助大数据分析技术从细微角度分析平台与互补者的竞争关系将是必要的。

2.2 产品同质化水平及测量

2.2.1 产品同质化水平的概念

产品同质化水平可以被解释为在一定范围内或一定时期内，市场中的同类产品不同企业或不同品牌的产品在外观、功能或营销方法等方面出现的相互模仿以至逐渐趋于相同的情况。在市场营销领域，对产品同质化的研究包括产品属性、产品类别和品牌概念等方面。例如 Park 在产品属性相似性的研究中，从产品特征的角度定义了相似性，其探究了产品特征相似性对品牌概念延伸的价值，Park 认为产品特征具有越高的相似性，品牌延伸的评价就越高^[20]。Muthukrishnan 在品牌延伸价值中将产品相似性相似性分为两种类型，一种是基于表层因素，另一种是基于深层因素。Bangyong 等通过产品的特征来分析不同页面的相似性。它们利用网页的标注信息和本体概念对网页进行分类^[21]。Zhai 等提出了一种新的从评论中提取产品特征并聚类相似产品特征的学习方法^[22]。Chang 等探讨了利益重叠和产品类别相似度对品牌延伸评价的差异。其研究发现类别相似度扩展更受关注预防的消费者青睐^[23]。

在平台生态中，基于产品同质化水平的研究主要包括电商平台的企业间或产品间的创新或模仿行为等课题。例如，夏季利用 TF-IDF 对产品文本进行向量化表达，使用余弦相似度计算了 Steam 平台中游戏产品描述的文本相似度。其探究了产品差异化定位对互补者绩效的影响^[24]。Huang 在淘宝平台的研究中发现，产品描述越相似的替代性产品会加剧市场的竞争程度，对焦点产品的销量产生消极影响^[6]。

2.2.2 产品同质化水平的测量

电商平台中的产品大多以文本、图象或者二者的混合形式进行展示。历史研究中提供了许多计算产品同质化水平或产品相似度的算法，包括文本和图像的算法。产品间的相似度计算方法与任何实体之间的相似性计算都大体相同，基本思

路都是将实体转化为数值向量从而进行计算。Colucci 等人评估了与电影相关的物品-物品相似度的感知正确性。其认为余弦相似度和 TF-IDF 算法在相似性计算中效果最好^[25]。在相似产品推荐的研究领域，Zhou 等人提出了基于图像提取的产品颜色和文本属性的相似度度量方法。他们通过考虑产品的颜色、属性和价格等特征来计算相似度^[26]。Huang 构建了产品相似网络来衡量产品相似性。因此，对于文本数据的相似度计算，过去研究更多还是基于 TF-IDF 或余弦相似度的方法，将产品文本进行直接数值量化然后计算。很少有研究从细粒度角度衡量产品相似度特征，把产品当作实体来研究其同质化程度。因此，将产品赋予一定的特征或属性，来描述其在客观世界中的存在将有助于对产品同质化水平的更细层次的研究，从而挖掘出更有价值的产品信息。

第三章 研究方法

本章介绍了本文在研究过程中使用到的技术方法，包括四个部分：命名实体识别、图结构与节点相似度计算、XGBoost 模型和机器学习模型的可解释性方法。其中：（1）命名实体识别技术用于抽取旅游产品文本标题数据中的各维度信息；（2）图结构的构建是基于命名实体识别技术抽取出的文本内容、在构建好的图网络基础上使用图计算方法计算出节点间相似性；（3）XGBoost 模型用于构建互补者的销量趋势预测模型；（4）基于 SHAP 的机器学习模型可解释性分析方法用于互补者销量趋势预测模型特征值的可解释性分析。

3.1 命名实体识别

3.1.1 命名实体识别的定义

命名实体识别（Name Entity Recognition, NER）的概念于 MUC-6（the Sixth Message Understanding Conference）第一次被提出^[27-28]。NER 是人工智能领域中自然语言处理（Natural Language Processing, NLP）领域中的一个子任务。其主要作用为从文本中识别出实体（例如组织、人员、地点等专有名词）以及含有特殊意义的时间、货币等数量短语并将其归类。举例说明：“美国商会会长苏珊在当地时间 1 月 12 日举行的美国商业状况年度演讲中称，竞争对美国的未来至关重要”。这句话中包含了三种实体，分别是：日期实体“1 月 12 日”、组织名称实体“美国商会”、姓名实体“苏珊”。

命名实体识别技术最早主要被用于识别某些特殊名词（实体、时间、数量）。随着研究的推进，学术领域开始关注对于开放域（Open Domain）的信息抽取研究，命名实体识别的任务不再局限于一般的实体分类，越来越多的实体类型被提出，研究学者们对其进行了更详细的任务划分：除了一般的实体识别之外，还可以针对某些特定领域（例如金融、生物、电影领域等等）进行专门的信息抽取。

随着互联网技术的发展以及大数据时代的来临，海量非结构化的互联网文本数据中蕴含着大量有价值信息，命名实体技术是进行文本挖掘与文本分析的核心技术，可以对数据进行有效的信息查找以及信息抽取。命名实体识别技术主要有三种典型的发展阶段。其分别为最开始基于规则和字典的方法、基于传统机器学习的方法以及目前被大规模使用的基于深度学习的方法。NER 在自然语言处理领

域中有着广泛的应用。例如构建知识图谱（Knowledge Graph）^[29]、机器翻译（Machine Translation）^[30]、网络搜索（Web Search）^[31]等。

3.1.2 命名实体识别的实现技术

命名实体识别技术主要有三种实现阶段，这三种实现阶段分别是：

（1）基于规则和字典的模式匹配方法

基于规则和字典的模式匹配方法是命名实体识别任务最初实现使用的技术。这种方法多需要语言学家的辅助，通过人工方式，按照数据集本身的结构特点制定针对性的规则模板或词典。这种模式匹配方法虽然准确性较高，但许多实体识别的规则构建依赖于特定的领域专家，且领域之间几乎没有重用。因此，该方法可拓展性低，需要消耗大量人力。

（2）基于传统机器学习的方法

命名实体识别任务在传统的机器学习方法中一般被作为序列标注问题实现。序列标记与传统的分类任务略有不同。对于序列标注，模型输出的预测标签一方面与当前输入的特征相关；另一方面，也与之前的预测标签相关。这表明预测标签序列之间存在很强的相互依赖性。传统机器学习中包括了隐马尔可夫模型（Hidden Markov Model, HMM）、最大熵马尔可夫模型（Maximum Entropy Markov Model, MEMM）、支持向量机 Support Vector Machine, SVM）等。这些模型都是常用的命名实体识别方法。

（3）基于深度学习的方法

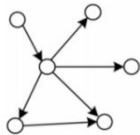
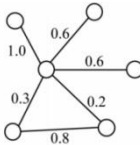
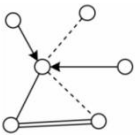
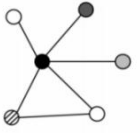
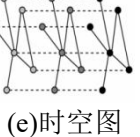
随着深度学习技术的发展，高维数据的处理成为其优势，命名实体识别的研究重点也开始转向深层神经网络（Deep Neural Network, DNN），该技术几乎不需要特征工程和领域知识^[33-35]，只基于标注好的训练数据，深层神经网络可以训练出达到最优目标的模型。基于深度神经网络的方法最早由 Collobert 等^[36]提出。之后，Ma 提出了一种 LSTM 与 CRF（条件随机场）相结合的 BiLSTM-CNNs-CRF 体系结构^[37]。该体系结构借鉴了 LSTM（长短期记忆神经网络）在自动分词任务上的优势，得到了更好的效果。Chiu 提出了一种双向的 LSTM-CNNs 架构。该架构可以自动检测到单词和字符级别的特征^[38]。Liu 等^[39]提出基于 LM-LSTM-CRF 的神经网络语言模型。在其研究中，文字识别神经网络语言模型被整合到多任务框架中，以提取文字水平的向量化表达。这些模型具有从大量数据中自动学习特征的功能。它们具有可扩展强且不依赖于域知识的优势。此外，还有其他深度学习方法，例如混合神经网络等。这些方法也被成功应用于命名实体识别工作中，均得到了理想的结果。

3.2 图结构与节点相似度计算

3.2.1 图结构定义与构建

图结构可以理解作为一种特殊的存储结构，主要由节点、边和权重构成。图网络的结构主要包括五种，分别是有向图、权重图、边信息图、节点异构图和时空图。其定义和示例图如表 3.1 所示^[41]。本研究基于命名实体识别技术抽取出了产品文本的相关实体，以产品的十个维度作为边的信息，以每个供应商所含有的每个实体的数量为权重构建了供应商-产品的图网络结构。图结构的构建在选定图结构的基础上，使用处理好的文本内容进行构建。在图结构的构建过程中，首先需要定义图节点、边信息和权重。在第四章中，本研究中使用命名实体识别技术抽取出来的实体内容作为图节点，产品维度信息作为边信息，每个供应商所含有的每个实体的数量为权重来构建图网络。

表 3.1 五种图结构的定义及示例²

图结构类型	定义	示例图
有向图	图结构中连接节点之间的边包含指向性关系，即节点之间的关联包含了方向的传递关系。	 <p>(a) 有向图</p>
权重图	图结构中的边包含权重信息，可以有效描述节点之间相互作用的可靠程度，定量表示关系的连接程度。	 <p>(b) 权重图</p>
边信息图	对于存在不同结构边的图结构，节点之间的关联关系可以包含权重、方向以及异构的关系，如在一个复杂的社交网络图中，节点之间的关联关系既可以是单向的关注关系，也可以是双向的朋友关系。	 <p>(c) 边信息图</p>
节点异构图	在图 G 中的节点属于多个不同类型的图结构，这种图结构通常根据异构节点类型对节点进行向量表示，也可以通过独热编码等编码方式实现节点的向量表示。	 <p>(d) 节点异构图</p>
时空图	时空图是一种属性图结构，其特点是高维特征空间 f^* 中的特征矩阵 X 会随时间变化，本文将其定义为 $G^* = (V, E, A, X)$ 。	 <p>(e) 时空图</p>

² 本表所有图片来自：王健宗, 孔令炜, 黄章成, 等. 图神经网络综述 [J]. 计算机工程, 2021, 47 (4):1-12.

3.2.2 节点相似度算法

节点相似度算法通常用于计算图结构中两个节点的相似度，用于计算相似度以及图结构中节点相似度的算法有很多，本研究中主要使用了 Adamic Adar 算法和修正的余弦相似度算法计算节点相似度。Adamic Adar 算法是公共邻居算法的改进版本。修正的余弦相似度算法是余弦相似度算法的改进版本。其定义和计算方法如下所示：

(1) 公共邻居算法

公共邻居算法在推荐领域使用广泛，通常被用作社区推荐算法。公共邻居捕捉到两个共同朋友比没有任何共同朋友的陌生人更有可能被介绍的想法。

该算法公式如下：

$$CN(x, y) = |N(x) \cap N(y)| \quad (3-1)$$

其中 $N(x)$ 是与节点 x 相邻的节点集， $N(y)$ 是与 y 节点相邻的节点集， $CN(x, y)$ 即为计算 x 和 y 节点的共同节点。值 0 表示两个节点不接近，而较高的值表示节点相近。

(2) Adamic Adar 算法

Adamic Adar 算法在公共邻居算法的基础上加入了节点的权重，该权重表示两节点间某邻居节点所拥有的邻居节点越多，该邻居节点在相似度计算中占据的权重越小。

该算法公式如下：

$$A(x, y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log |N(u)|} \quad (3-2)$$

其中 $N(u)$ 是与相邻的节点集 u 。值 $A(x, y) = 0$ 表示两个节点不接近，而较高的值表示节点较近。Adamic Adar 相似度算法是在共同邻居算法的基础上加入了衡量邻居节点受欢迎程度的权重，在计算两节点单纯的公共邻居节点个数的同时，加入了全局特征。

(3) 修正的余弦相似度算法

修正的余弦相似度与一般所说的余弦相似度略有不同，其计算公式如下：

$$\text{sim}(i, j) = \frac{\sum_{u \in U} (R_{u,i} - \bar{R}_u)(R_{u,j} - \bar{R}_u)}{\sqrt{\sum_{u \in U} (R_{u,i} - \bar{R}_u)^2} \sqrt{\sum_{u \in U} (R_{u,j} - \bar{R}_u)^2}} \quad (3-3)$$

修正的余弦相似度的提出是为了将向量各维度的量纲差异性考虑在内，从而

解决余弦相似度算法仅考虑向量维度方向上的相似性的问题。因此修正的余弦相似度算法在计算相似度的时候，在每个维度上减去了均值。

3.3 XGBoost 模型

3.3.1 XGBoost 简介

XGBoost 基于 boosting 算法的迭代决策树思想提出^[50]。该模型梯度提升树 (GBDT) 的改进，属于集成学习模型的一种，可以用来做回归和分类任务。XGBoost 的优点在于能够自动利用 CPU 的多线程进行并行，同时在算法上加以改进提高了精度。XGBoost 算法在许多预测任务中都有良好的表现，许多历史研究也都基于该模型并取得的良好预测效果。XGBoost 已经成为了各式 AI 竞赛中出场率最高的算法之一。XGBoost 具有许多优点。其在高维特征数据中表现良好^[42]。可以选择含有最重要信息成分的数据进行预测，灵活处理预测因子之间存在潜在的高纬度相关性。

3.3.2 XGBoost 模型原理

XGBoost 算法的基本思想是基于训练集得到 k 个分类器。每棵树在预测后，每个样本都会有一个分数。当需要对给定的预测样本进行预测时，由每棵树的预测概率和每棵树的预测分数的综合作用得到该样本的预测结果。其预测结果定义为：

$$\hat{y} = \sum_{k=1}^K f_k(x_i), f_k \in F \quad (3-4)$$

其中， F 是所有树的集合， \hat{y} 表示对样本的预测结果， $f_k(x_i)$ 表示 k 个树分类器对样本 x_i 的预测评分。由此，模型目标函数定义为：

$$Obj(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (3-5)$$

目标函数是由两个部分构成的，分别是误差函数项和正则化项。其中，误差函数用于逐步优化目标函数。误差函数首先优化第一棵树，优化完成之后再优化第二棵树，直至优化完 K 棵树，过程如下：

$$\begin{aligned}
\hat{y}_1^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\dots\dots\dots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned} \tag{3-6}$$

其中, $\hat{y}_i^{(t)}$ 表示模型的最终预测结果, $\hat{y}_i^{(t-1)}$ 表示前 $t-1$ 轮的模型预测结果, $f_t(x_i)$ 表示当前新加入的树模型。

$$obj(\theta)^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \sum_{k=1}^K \Omega(f_k) \tag{3-7}$$

将 (3-6) 式中的 $\hat{y}_i^{(t)}$ 代入 (3-7) 式中可以得到:

$$obj(\theta)^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + C \tag{3-8}$$

当 $t-1$ 轮确定后, 之前 $t-1$ 棵树的复杂度为 常值 C 。则目标函数可以表示为:

$$f(x + \Delta x) \cong f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2 \tag{3-9}$$

对于目标函数进行泰勒展开:

$$f(x + \Delta x) \cong f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2 \tag{3-10}$$

(3-10) 式去掉常数项, 进行泰勒展开后的目标函数为:

$$obj(\theta)^{(t)} \cong \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) + \Omega(f_i) \tag{3-11}$$

其中, 一阶导数 $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$, 二阶导数 $h_i = \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)})$ 。针对第二部分的模型复杂度, 将叶子结点权重和树的深度考虑到目标函数的正则化中:

$$\begin{aligned}
obj(\theta)^{(t)} &\cong \sum_{i=1}^n [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_i) \\
obj(\theta)^{(t)} &\cong \sum_{i=1}^n [g_i w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^t w_j^2 \\
obj(\theta)^{(t)} &\cong \sum_{i=1}^n [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T
\end{aligned} \tag{3-12}$$

此时的目标函数就是求解一元二次方程的过程。

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

$$obj^* = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} \quad (3-13)$$

根据得分函数可以定义分割的增益函数为：

$$Gain = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} - \lambda \quad (3-14)$$

式（3-14）代表左侧子树分数与右侧子树分数的和减去未分割的时候的分数。 λ 为新加入叶结点引入的参数。该参数表示复杂度的代价。 $Gain$ 表示分裂时的增益得分。该得分为正值且数值越大时，表示该结点拥有较大的切分价值，即切分后 obj^* 减小。同理， $Gain$ 为负值则表示该节点的切分价值较小，即切分后 obj^* 增加。 γ 值越大时，代表切分后对 obj^* 的下降程度有更严苛的约束。该过程描述了 XGBoost 模型如何通过枚举得到树的最优结构。

3.4 机器学习模型的可解释性方法

3.4.1 可解释性理论

在机器学习领域，人们通过数据来训练模型以达成良好的预测和分类任务已经得心应手。在大数据背景下，线性回归模型不一定可以永远很好地拟合所有数据。近些年来，关于可解释性机器学习的研究也愈发普遍，不管是传统的机器学习模型还是深度学习模型，人们不再在精确度和可解释性之前寻求平衡，而是希望在获得高准确度的模型结果同时也了解模型是如何基于数据做出预测的。

机器学习模型的作用原理本质上是依据计算机的大规模计算能力，通过训练数据来寻找数据之间的隐含规律，从而得到一个函数来拟合样本得到预测结果。模型的预测就是将输入的多维特征映射成预测结果，模型的可解释性则是对这个函数进行解释，即解释该函数的映射结果，如何映射出结果，以及为什么映射出该结果。解释函数的难易程度则被认为是机器学习模型可解释性高低的一种表现。目前，关于模型的可解释性理论主要研究三个方面的问题：（1）单个样本角度，研究每个特征如何影响模型预测结果；（2）预测结果角度，研究机器学习模型中特征的重要性程度；（3）整体样本角度，研究每个特征如何影响模型预测。

3.4.2 基于 SHAP 的可解释分析方法

关于模型可解释性的方法有很多，目前基于 SHAP 的模型可解释分析方法被广泛使用，这也是本研究采用的模型可解释方法。SHAP 由 Lundberg 等人于 2017 年提出^[43]。该方法统一了六种可加特征归因方法，这六种方法分别是：LIME、DeepLIFT、分层相关传播、Shapley 回归值、Shapley 采样值和定量输入影响^[44-49]。这种解释预测框架将模型的预测结果解释为由每个输入特征的归因值之和。

Lundbrg 指出，一个简单模型最好的解释方法则是模型本身，对于复杂模型来说，并不能使用原始模型本身来理解，我们必须使用一个简单的解释模型，将这个简单的解释模型定义为原始模型的任意解释逼近。这一思想简单直白地表面了模型可解释性的宏观指导原理。SHAP 将 Shapley 值解释表示为了一种可加特征归因的方法，SHAP 将模型预测值解释为每个输入特征的归因值之和：

$$g(z') = \phi_0 + \sum_{j=1}^M \phi_j z'_j \quad (3-15)$$

(3-15) 式中， g 代表解释模型， $z' \in \{0,1\}^M$ 代表相应特征能否能被观察到， M 表示输入的特征数量， $\phi_j \in \mathbb{R}$ 则表示各特征归因值， ϕ_0 是解释模型的常数项。由于树模型的输入必须是结构化数据，对于实例 x ， z' 应该是所有值为 1 的向量，即所有特征都能被观察到的，于是该公式可以简化为：

$$g(x') = \phi_0 + \sum_{j=1}^M \phi_j \quad (3-16)$$

SHAP 有三个优良的特征属性：（1）局部准确性，即特征归因的总和等于需要解释的模型的输出；（2）缺失性，即表示缺失特征的归因值为零；（3）一致性，即若模型发生了改变，使特征值边际贡献增加或维持不变（与其他特征无关），则归因值也会增加或维持不变。

第四章 数据与变量测量

本章内容主要分为两大部分：产品同质化水平指标的计算和销量趋势预测模型训练所需的变量介绍。4.1 节介绍了本研究的主要数据来源；4.2 节阐述了产品相似度提取和相似度计算的具体过程，包括旅游产品特征维度的定义与提取、供应商-产品网络的构建和产品同质化水平的计算；4.3 节介绍了本研究中用于训练销量趋势预测模型所使用的全部变量。图 4.1 展示了本研究的所有变量组成。

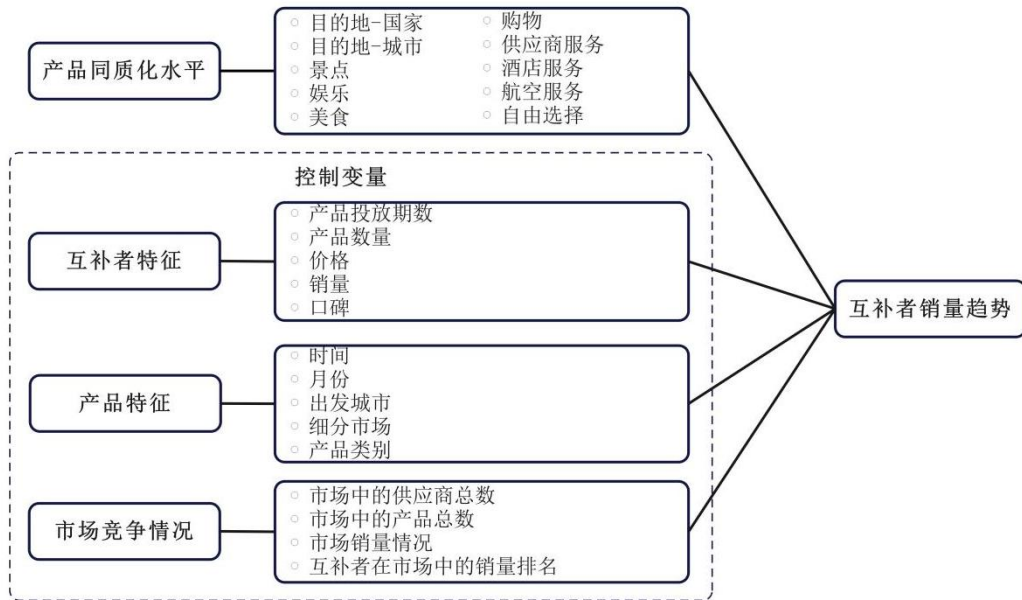


图 4.1 本研究的所有变量组成

4.1 数据

本研究的实验数据来自于携程平台（<https://hotels.ctrip.com>）。携程是一家领先的在线旅行平台，提供广泛的旅游服务，包括机票预订、酒店预订、度假套餐和其他目的地服务。该平台覆盖全球，提供 200 多个国家和地区的 140 多万家住宿和 200 多万种旅游产品。截至 2021 年，携程的市场份额约为 70%。由于 2020 年农历春节假期前出现的新冠疫情，中国出境旅游市场的繁荣指数开始大幅下降。为避免疫情爆发对出境旅游的影响，本研究选取了 2017 年 3 月至 2018 年 12 月期间的美国出境游的全部数据。本文在每个月第一天进行数据爬取数据，总共爬取

了 42,252 个产品信息条目。我们爬取了产品 id、标题、服务保障、供应商、产品卖点、价格、销售量、评论量、评分等信息。图 4.2 显示了携程网站上一个目的地是美国的度假产品的信息。

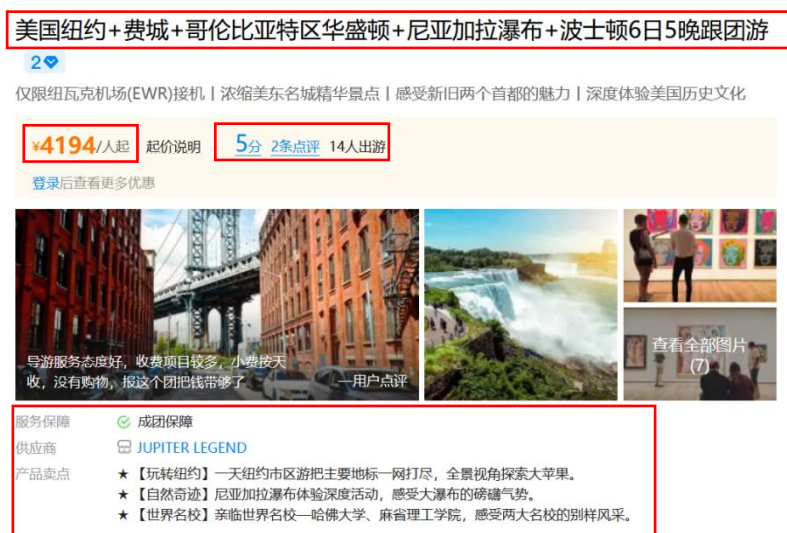


图 4.2 携程平台上的产品页面

4.2 产品特征提取和相似度计算

为了探究与平台产品同质化程度是否可以允许互补者对自己未来的销量进行预测, 合适地衡量互补者与平台产品的同质化程度是必要的。介于本研究的层面是供应商水平而非产品水平, 本研究面临的挑战是: (1) 如何从产品角度来表示并提取供应商的产品投放特征? (2) 如何计算互补者与平台产品的同质化程度? 4.2.1 节和 4.2.2 节阐述了本研究从产品角度来表示供应商产品投放特征的具体过程, 包括将旅游电商产品分解为九个子维度以及使用自然语言处理技术进行实现。4.2.3 节则展示了本研究通过构建图网络计以及图计算来计算互补者与携程平台之间的产品同质化程度的具体细节。

4.2.1 旅游产品特征维度定义

在旅游公司向平台投放自身产品的过程中, 大部分平台互补者的业务相对平台来说更为单一和集中——这并不难理解, 因为众多的小型互补者所拥有的旅游资源是有限的, 所以供应商只能基于自身所能提供的产品内容来推出相应的旅游产品。因此, 这种很难更改的自身资源属性则反馈在供应商所投放的旅游产品内

容中。

为了更好的解释，下面将展示一个简单的例子。将美国出境游前五家头部公司（按照市场份额排序）在 2017 年 12 月的所有产品目的地城市的内容提取出来以代表其产品的目的地维度的内容（具体数值在表 4.1 中展示），然后进行量化并将它们映射到二维平面上（如图 4.2）。需要强调的是，反映供应商的目的地维度的数据维度总共有 127 维，实验中使用 PCA 降维将其降到了 3 维，分别反映到了气泡图的横纵坐标轴和大小的层面。因此这意味着气泡图的横纵坐标并不具有强烈的实际意义，但仍可以看出，每个公司在坐标轴平面上的分布是不同的，距离越近两家的公司代表他们的产品同质化水平越高。例如，从表 4.1 中可以看出成都途风国际旅行社有限公司在我们展示出的六个目的地内容的产品数量与携程平台具有大致相似的趋势，因此这两者在气泡图中的位置也更为接近。在一定程度上，这直观体现了每个公司侧重于不同数量的目的地城市的产品投放。

表 4.1 2017 年 12 月美国出境游销量排名前五家头部公司产品的目的地内容计数(个)

供应商	华盛顿	费城	纽约	拉斯维加斯	洛杉矶	夏威夷
上海携程国际旅行社有限公司	71	73	69	58	51	21
成都途风国际旅行社有限公司	69	61	76	88	81	5
北京纳美旅行社有限公司	40	33	36	50	53	30
北京达美国国际旅行社有限责任公司	22	18	30	19	20	39
Messi International Travel LLC	0	0	0	80	79	0

注:表中只列举了六个目的地城市。实际上用于计算的目的地城市共有 127 个。

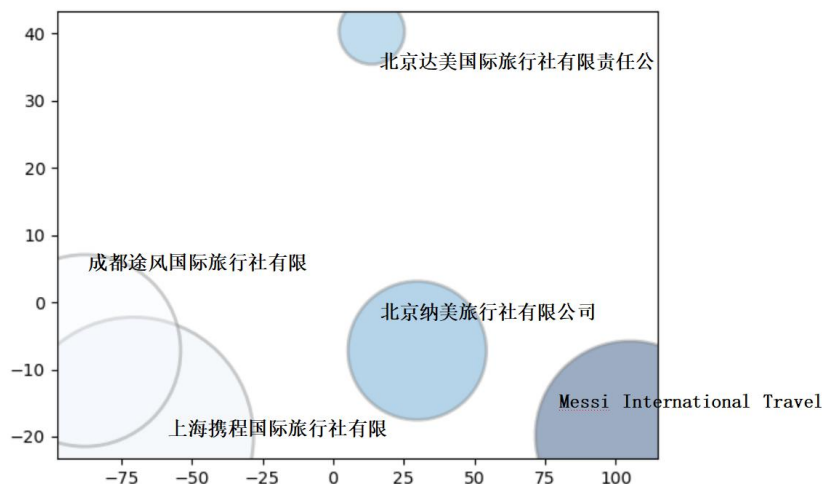


图 4.3 2017 年 12 月美国出境游销量排名前五家头部公司产品的目的地内容投放偏好可视化

注：横纵坐标不具有实际意义

上文简单介绍了本文如何从产品角度来体现供应商的产品投放偏好并量化供应商间产品相似度的例子。旅游产品作为一款体验类产品，除了目的地城市之外还包括许多其他内容维度，如何将旅游产品的结构化表征对本文更全面地表示供

应商产品投放偏好将是有帮助的。然而，旅游产品作为一款体验类产品，很少有研究将其进行概念化表示。1994 年 Smith 将市场营销和供应方的观点进行融合，首次将旅游产品分解为了五个元素，即：物理实体、服务、好客程度、选择自由度和游客参与程度^[7]。Xu 则基于测试，检验了前者的概念化结构，在 2010 年基于新视角的基础上对其进行了改进，将后四个元素视为平等关系（服务、好客程度、选择自由度和游客参与程度）而非 Smith 提到的包含关系^[8]。因此，我们根据过去提出的构成旅游产品的五大维度，定义了针对旅游电商产品的分析维度，其详细信息展示在了表 4.2。

表 4.2 旅游电商产品的分析维度

元素	解释	子维度
the physical plant	场地、自然资源、设施、酒店、邮轮、天气、水质、拥挤程度、旅游基础设施（瀑布、野生动物、度假胜地）	目的地（国家、城市） 景点 美食 娱乐 购物
service	酒店管理，机场服务，旅行社服务	旅行社服务（司机、导游、接送机等） 机场服务（直飞、航空公司类别等） 酒店服务（酒店的星级、民宿类别等）
hospitality	旅游目的地、酒店等人员热情，好客程度	/
freedom of choice	自由购物、自由选择航空公司、汽车路线、酒店或餐厅	自由选择（自由活动、自驾、自由购物等）
involvement	旅客个人在旅行中获得的主观感受	/

以下是对 Smith 提出的五大旅游产品结构的解释及本研究相应的调整说明：

（1）**the physical plant**：这一元素被解释为场地、自然资源、设施、酒店、邮轮、天气、水质、拥挤程度、旅游基础设施（瀑布、野生动物、度假胜地）等物理实体。我们定义了目的地（包括国家、城市）、景点、美食、娱乐、购物这五个子维度来反映旅游产品中该元素的相关信息；

（2）**service**：本研究定义了旅行社服务、机场服务、酒店服务这三个子维度来反应旅游产品中与“服务”相关的信息。

（3）**hospitality**：这一元素被定义为好客程度，例如目的地的一些景点和入驻酒店等员工区别于高效完成工作的热情程度。但在电商产品文本信息中很难得到对这一元素直观的反映，因此这一元素被我们省略。

（4）**freedom of choice**：本研究定义了“自由选择”这一子维度来反映该元素。“自由选择”在产品文本中出现的主要内容有：自由活动时长、自驾、自由购物等。

(5) **involvement**: 这一元素被解释为旅客个人在旅行中获得的主观感受。介于这是一个偏向主观的元素, 本研究使用了产品的评分来反应这一内容, 认为是否给予了游客积极体验感的产品将反映在产品评分中, 该变量后期被纳入互补者特征的变量中。

4.2.2 产品特征抽取

根据 4.2.1 节中定义的九大旅游产品子维度, 本节使用 BIO 标注原则标注了 2000 条产品文本数据, 并使用了 BERT+BiLSTM+CRF 深度学习模型进行训练和抽取, 模型的抽取结果示例表 4.3 中展示。

由于抽取出的文本数据中仍存在的一些噪声数据可能会影响我们进一步的相似度计算结果, 实验对抽取出的实体进行了进一步的数据清洗, 具体操作包括:

(1) 实体删除: 删除特殊符号、删除空值、删除出现频率小于 10 的实体 (经过统计分析实体频率小于 10 的实体没有实际意义、删除字符长度为 1 的实体。

(2) 实体对齐: 合并表达相同语义的实体 (例如“四星级酒店”和“四星酒店”)。

关于使用命名实体识别技术进行文本抽取的技术细节在附录中展示。

表 4.3 模型抽取结果示例

产品文本	美国旧金山+拉斯维加斯+洛杉矶·12 日 10 晚半自助游, 机场直飞免费托运, 全程四星酒店·一号公路自驾+金门大桥+游船+2 日自由活动+奥特莱斯+牛排餐·全程中文导游服务+机场接送
每个维度下抽取结果:	
Destination-country	美国
Destination-city	旧金山; 拉斯维加斯; 洛杉矶
Attraction	一号公路; 金门大桥
Food	牛排餐
Entertainment	游船
Shopping	奥特莱斯
Provider_service	中文导游服务; 机场接送
Airline_service	直飞; 免费托运
Hotel_service	四星酒店
Freedom_choice	2 日自由活动; 自驾

4.2.3 互补者与平台间的产品同质化水平计算

供应商投放的产品内容将以拆分的这些子维度来结构化表示。因此, 可以使用这些子维度表示每个供应商在某一时期的产品投放特征 (公式如下)。

$$C_{it} = F(X_{it}^{Destination}, X_{it}^{Attraction}, X_{it}^{Service}, X_{it}^{Food}, X_{it}^{Entertainment}, X_{it}^{Shopping}, X_{it}^{FreedomChoice}) \quad (4-1)$$

下一步需要考虑的问题则是寻找一个合适的计算方法来衡量互补 *jianli* 者与平台间的产品同质化程度，即数值水平上的相似度。排除最后一期产品数据，对于每个时期都构建了一个图网络来储存这些供应商产品投放特征的信息，每个时期的每个细分市场和每个产品类别也分别构建图网络。具体来说，图由两种节点和赋有权重的有向边组成，有向边的权是每个子维度内容在供应商自身产品中所出现的比重。即 $G = \{(m, r, n), w\} \subseteq E^{provider} \times R \times E^{entity} \times W$ ， $E^{provider}$ 是所有供应商的集合， R 是节点间关系的集合， E^{entity} 是产品子维度内容的集合。 $m \in E^{provider}, n \in E^{entity}, r \in R, w$ 是有向边的权重， $w \in W$ 。例如，计算供应商 i 在 t 时期时指向代表“纽约”的内容实体的计算方法为：

$$w(h_{it} \rightarrow n) = \frac{totalProductNum_NewYork_{it}}{totalProductNum_{it}} \times 100\% \quad (4-2)$$

其中， $w(h_{it} \rightarrow n)$ 表示图网络中供应商 i 在 t 时期内指向子维度内容实体 n 的权重； $totalProductNum_{it}$ 表示供应商 i 在 t 时期的产品总数； $totalProductNum_NewYork_{it}$ 表示供应商 i 在 t 时期拥有目的地城市为“纽约”的产品总数。图 4.4 展示了关于图网络的模式图。供应商-产品网络的 Neo4j 可视化结果在附录 B 中展示。

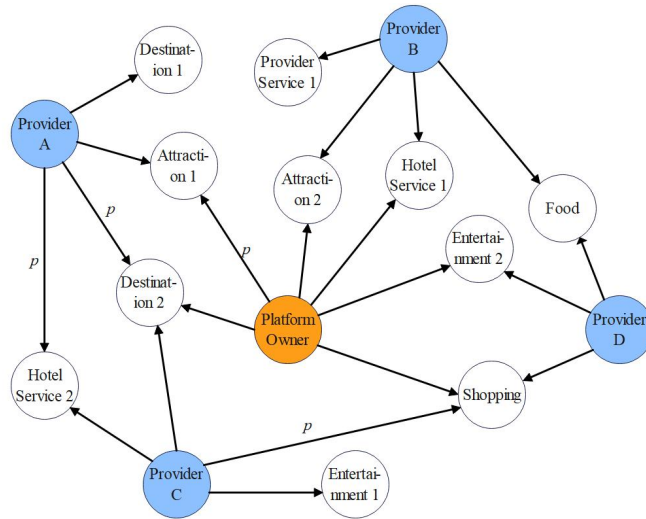


图 4.4 t 时期下，供应商及其产品投放内容的图网络结构

在构建好的图网络基础上计算出互补者与携程的节点间相似度即可表示在 t 时期下，互补者与携程平台产品同质化程度。由于任意两个供应商之间并没有权

重边直接相连，而是共同指向子维度的内容实体；同时考虑到图网络中增加了权重，因此本研究选择了在社区推荐算法中常用的 Adamic Adar 算法和修正的余弦相似度算法来衡量节点的相似性。Adamic Adar 相似度的计算方法是共同邻居相似度（Common Neighbours Similarity）的改进。

$$CN(x, y) = |N(x) \cap N(y)| = \sum_{u \in N(x) \cap N(y)} N(u) \quad (4-3)$$

如公式（4-3）所示，共同邻居相似度算法的思想很简单：对于图网络 $G = (V, E)$ 中的两个节点 $x, y \in V$ ，如果这两个节点的共同邻居数越多，表示其相似度越高，反正则越低。使用共同邻居算法求出的相似度数值的实际意义则是 x, y 所拥有的公共邻居节点的个数。

$$A(x, y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\ln |N(u)|} \quad (4-4)$$

Adamic Adar 相似度算法则是在共同邻居算法的基础上加入了衡量邻居节点受欢迎程度的权重，如公式（4-4）所示。权重项 $\frac{1}{\ln |N(u)|}$ 表示： x, y 所共有的某个邻居节点 u 相比其他邻居节点连接到了除 x, y 外的更多邻居节点时，该节点 u 的权重就会被降低。接着，加入 x, y 两节点连接共同邻居节点的权重向量 W_x, W_y 的余弦值来修正 Adamic Adar 相似度所得出的相似度。这种做法是为了体现出同一种产品内容在不同供应商中的占比。例如，供应商 A 与携程平台拥有完全相同的邻居节点 u_1 和 u_2 ，这两个节点各自拥有的邻居节点数量也相等。在这种情况下，使用 Adamic Adar 相似度将得到相同的值。然而假设供应商 A 的所有产品中含有 u_1 和 u_2 的比重分别是 0.5 和 0.9，而携程是 0.1 和 0.2——显然两家公司的产品投放内容是有差异的。因此，加入修正的余弦相似度将可以避免这一问题。最后，由于修正的余弦相似度的值域为 $(-1, 1)$ ，本文将其映射到了 $(0, 1)$ 之间。本研究最终得到的进行图节点相似度计算的具体方法如下所示。

$$S(x, y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\ln |N(u)|} \times \frac{CS(W_x, W_y) + 1}{2} \quad (4-5)$$

4.3 变量定义

4.2 节重点阐述了本文是如何从产品的非结构化文本数据中量化得到供应商间产品同质化程度这一指标的。由于本研究重点关注该指标是否对供应商未来的销量趋势的预测起重要作用，因此本节加入了其他变量以全面描述特定时期内的供应商特征，包括：互补者特征、市场竞争情况和产品特征。本节将描述这些变

量的定义及处理方式。

4.3.1 产品特征

携程平台上的旅行产品种类繁多，因此本文考虑了一些产品特征变量来控制实验，这些产品特征变量如下：

（1）时间

本研究的数据是从 2017 年 3 月至 2018 年 12 月的时间序列数据，由于本研究的目的是预测销量趋势（即下一期供应商的销量是上升还是下降），所以时间作为一个很重要的宏观变量纳入了考虑范围。

（2）月份

由于考虑到旅游产品的销量受月份、季节因素影响很大，本文也将月份变量纳入考虑因素中。月份变量是区别于时间变量的，它不会体现出年份的变化。

（3）出发城市

旅游产品通常包含一个国内出发地城市，旅游市场的差异也会对产品销量产生影响。然而，出发城市并不属于本研究的重点研究变量。考虑到北京的旅行市场规模庞大、产品多元丰富，在后文的实际实验过程中，本文全部选择了出发城市为北京的旅游产品。

（4）细分市场

本研究的所关注的重点变量是产品同质化水平这一指标，即互补者与携程平台之间的产品文本相似度的变化是否会对互补者产生影响。在 4.2 节中我们描述了如何从产品的微观层面映射到供应商的想法。我们考虑到，供应商在面对不同的旅游目的地提供的服务内容通常是不同的，例如，旅游目的地为美东地区的产品会推出一些名校游览的项目，而美西地区的娱乐项目更多，在这种情况下，对产品范围进行约束是必要的。本文按照携程平台上为每个产品展示的“目的地城市”这一信息作为细分市场变量，数据共包含 42258 个产品，本研究共研究了四个细分市场的产品，产品体量占据总体量的 70% 以上，选取的细分市场以及其在观测周期内的产品数量如表 4.5 所示。

（5）产品类别

除了细分市场之外，携程平台的产品通常分为一些具体的出行类别，一般有跟团游、半自助游、自由行、自家团、游学游等。不同出行类别的产品内容、价格都存在差异。因此，在控制了计算供应商之间产品相似度时的细分市场变量之外，还需要对产品类型进行限制，由于携程平台上不同细分市场下产品类型数量差异较大，本文只选择了跟团游、半自助游和自由行这三类产品，每个细分市场

下的不同类别的产品数量如表 4.5 所示。

表 4.5 观测周期内细分市场及其产品、供应商数量

		洛杉矶	纽约	塞班岛	旧金山
产品类别	总数	12568	8061	5699	4495
	跟团游	7590	3252	356	2944
	半自助游	3704	3388	3873	968
	自由行	411	1021	1389	237
	私家团	862	399	80	342
供应商数量	总数	146	148	129	124
	跟团游	131	131	38	103
	半自助游	35	68	107	26
	自由行	3	12	56	5
	私家团	15	12	8	21

4.3.2 互补者与平台间的产品同质化水平

互补者与平台间的产品同质化水平即为每个时期内每一个互补者在特定的细分市场和特定产品类别下与携程平台的产品同质化水平。本文基于图网络计算出了每个时期内每个互补者与平台的相似度数值。另外，由于本文的目的是预测互补者未来销量，变量中将包含本时期 t （非初始时期）内的相似度以及从初始时期至 $t-1$ 时期的累计平均相似度。关于产品相似度的指标及其定义在表 4.6 中展示，其统计数据汇总信息在附录 C 中展示。

表 4.6 变量定义：互补者与平台间的产品同质化水平

子类别	变量	定义
目的地城市	city_sim _{it}	t 时期内特定细分市场下供应商 i 所有旅游产品在景点维度的内容相对于平台的相似度
景点	attraction_sim _{it}	t 时期内特定细分市场下供应商 i 所有旅游产品在景点维度的内容相对于平台的相似度
酒店服务	hotel_service_sim _{it}	t 时期内特定细分市场下供应商 i 所有旅游产品在酒店维度的内容相对于平台的相似度
航空服务	airline_service_sim _{it}	t 时期内特定细分市场下供应商 i 所有旅游产品在航空服务维度的内容相对于平台的相似度
供应商附加服务	provider_service_sim _{it}	t 时期内特定细分市场下供应商 i 所有旅游产品在供应商附加服务维度的内容相对于平台的相似度
购物	shopping_sim _{it}	t 时期内特定细分市场下供应商 i 所有旅游产品在购物维度的内容相对于平台的相似度
娱乐	entertainment_sim _{it}	t 时期内特定细分市场下供应商 i 所有旅游产品在娱乐维度的内容相对于平台的相似度
美食	food_sim _{it}	t 时期内特定细分市场下供应商 i 所有旅游产品在美食维度的内容相对于平台的相似度
自由选择	freedom_choice_sim _{it}	t 时期内特定细分市场下供应商 i 所有旅游产品在自由选择维度的内容相对于平台的相似度

4.3.3 互补者特征

过去研究表明, 电商平台中的产品销量往往还与供应商自身与市场环境相关, 因此本研究从内部和外部角度提取了一些特征。本节将讨论内部角度, 4.3.4 节将讨论外部竞争因素。为了全面地描述互补者的特征, 本文加入了以下变量:

(1) 产品投放期数

过去研究表明, 越晚建立的商业体会比建立更长久的商业体更倾向于失败^[15]。在本研究中, 入驻平台较久的互补者将会在平台中更长期投放产品。因此, 本研究计算了每个供应商(包括携程平台本身)在观测期内所投放的产品期数, 观测数据内最大的产品期数则是观测时长。

(2) 产品数量

对于不同互补者来说, 由于其个体规模的不同, 与平台产品相似度的变化对其影响很可能是不同的。因此本文把互补者在某一特定时期内在特定细分市场 and 整个美国市场所投放的全部产品数量也分别纳入参考。另外, 本文还观察到对于互补者来说, 其全部销量在每个产品上的分布通常不是均匀的: 一些销量比较好的产品销售量可能会占总销量的绝大部分, 在这种情况下, 本文加入了销量不为 0 的产品总数作为变量; 最后, 由于本文在比较互补者与平台间产品相似度的时是基于相同细分市场来进行比较。因此, 本文关注在某一个特定时期的某一个特定细分市场下, 互补者在此细分市场的产品投放占比。本文假定对于专注与某个细分市场的互补者和专注多个细分市场的互补者受到与平台产品相似度变化的影响是不同的。

(3) 价格

许多历史研究表面, 产品的价格会显著影响销量, 因此本文将互补者在每个时期内的产品价格纳入参考, 包括产品本期平均价格和产品累计平均价格。同时, 价格也同样作为平台与互补者的竞争因素而决定着消费者的选择, 在此基础上, 还加入了每个时期互补者与携程平台的价格差作为参考。

(4) 销量

互补者历史时期的销量情况将会极大影响未来的销量趋势, 与产品数量的变量定义相似, 本文也加入了五个变量来表示互补者的销量特征, 包括特定细分市场和整体市场的当期和过去销量情况以及特定细分市场的销量占比。

(5) 口碑

网络口碑也是影响消费者消费决策的因素之一。过去的研究表明, 评论数量、评分等因素会影响消费者的消费决策从而对网络销量产生影响。因此在此将评论

数量和产品评分也作为变量。由于旅游产品的消费频率偏低，携程平台上的产品评论数量较少且大部分为默认评价，因此本文并没有将关于评论的情感分析等内容纳入参考。

关于供应商特征的变量及其定义在表 4.7 中展示，其统计汇总信息在附录 C 展示。

4.3.4 市场竞争情况

互补者与平台之间存在竞争关系，而市场上的来自其他互补者的竞争情况也决定着互补者的销量，因此本文考虑了以下变量来全面表示市场内的竞争情况：

（1）供应商总数

考虑到在同一市场内，供应商的总数越多，意味着市场竞争也越激烈。因此本文衡量了每个时期内，每个细分市场下所有的供应商总数。

（2）产品总数

与供应商总数同理，产品总数越多在一定意义上也意味着该市场内的竞争压力更大，因此本文衡量了每个时期内，每个细分市场下所有的产品数量。

（3）市场总体销量

旅游产品的销量和节日、季节等时间因素有着密切关系。在每个时期内，市场总体销量也会影响到互补者的绩效水平。因此本文衡量了市场总体销量作为变量之一。

（4）销量排名

销量排名是与互补者个体相关的，随着时间变化的变量，表示互补者在每个时期每个细分市场下的销售排名。该变量用于衡量当期销量的相对水平。

市场竞争相关的变量定义以及统计汇总信息在表 4.8 及附录 C 中展示。

表 4.7 变量定义：供应商特征

子类别	变量	定义
产品投放期数	<i>number_of_periods_{it}</i>	供应商 <i>i</i> 进入市场时截止到 <i>t</i> 时期的生存时间（最大时间为观测时长）
产品数量	<i>submarket_product_num_{it}</i>	供应商 <i>i</i> 在 <i>t</i> 时期内向某细分市场投放的所有产品数量（以产品 <i>id</i> 为标准）
	<i>submarket_product_withsales_num_{it}</i>	供应商 <i>i</i> 在 <i>t</i> 时期内向某细分市场投放的所有本期销量不为 0 的产品数量（以产品 <i>id</i> 为标准）
	<i>total_product_num_{it}</i>	供应商 <i>i</i> 在 <i>t</i> 时期内所有投放的产品数量（以产品 <i>id</i> 为标准）
	<i>total_product_withsales_num_{it}</i>	供应商 <i>i</i> 在 <i>t</i> 时期内所有投放的本期销量不为 0 产品数量（以产品 <i>id</i> 为标准）
	<i>product_num_percent_{age_{it}}</i>	供应商 <i>i</i> 在 <i>t</i> 时期内在某细分市场下所投放的产品数量占自身全部投放产品数量的比例（例如，供应商 A 在 <i>s1</i> 市场投放产品 100 个，在 <i>s2</i> 市场投放产品 300 个，则其在 A 市场的 <i>product_percentage</i> 为 25%）
产品价格	<i>price_avg_{it}</i>	供应商 <i>i</i> 在 <i>t</i> 时期的产品平均价格
	<i>price_avg_cum_{it}</i>	供应商 <i>i</i> 在 <i>t</i> 时期的过去累计产品的平均价格
	<i>price_gap</i>	供应商 <i>i</i> 在 <i>t</i> 时期与携程平台在 <i>t</i> 时期在某细分市场下产品平均价格的差
销量	<i>submarket_sales_{it}</i>	供应商 <i>i</i> 在 <i>t</i> 时期的某细分市场下的产品销量
	<i>submarket_sales_cum_{it}</i>	供应商 <i>i</i> 在 <i>t</i> 时期的某细分市场下的过去累计平均产品销量
	<i>total_sales_{it}</i>	供应商 <i>i</i> 在 <i>t</i> 时期的所有产品销量
	<i>total_sales_cum_{it}</i>	供应商 <i>i</i> 在 <i>t</i> 时期的所有产品的过去累计平均产品销量
	<i>product_sales_percent_{age_{it}}</i>	供应商 <i>i</i> 在 <i>t</i> 时期内在某细分市场下所投放的产品销量占自身全部投放产品销量的比例
口碑	<i>comments_num_{it}</i>	供应商 <i>i</i> 在 <i>t</i> 时期的产品总评论数量
	<i>comment_num_cum_{it}</i>	供应商 <i>i</i> 在 <i>t</i> 时期的过去平均产品总评论数量
	<i>product_score_{it}</i>	供应商 <i>i</i> 在 <i>t</i> 时期的产品评分均分（销量不为 0 的产品）
	<i>product_score_cum_{it}</i>	供应商 <i>i</i> 在 <i>t</i> 时期的累计过去产品评分均分（销量不为 0 的产品）

表 4.8 变量定义：市场竞争情况

子类别	变量	定义
供应商总数	<i>total_providers_t</i>	<i>t</i> 时期该细分市场下所有供应商的总数
	<i>total_providers_haveSales_t</i>	<i>t</i> 时期该细分市场下所有销量不为 0 的供应商总数
产品总数	<i>total_products_t</i>	<i>t</i> 时期该细分市场下所有的产品总数
	<i>total_products_haveSales_t</i>	<i>t</i> 时期该细分市场下所有销量不为 0 的产品总数
市场整体销量	<i>total_travle_count_t</i>	该细分市场下的市场整体销售量
销售排名	<i>sales_rank_{it}</i>	供应商 <i>i</i> 的产品总销量在所有供应商中的排名

第五章 销量趋势预测模型

5.1 数据集预处理

4.3 节已经详细介绍了本研究的变量定义。然而，在进行预测之前，为了让预测模型达到更好的效果，实验前还需要对数据进行进一步的处理以输入模型进行训练。第一，将时期变量 t 转换成的数值变量，从 2017 年 3 月至 2018 年 12 月共 22 期，所以变量 t 对应的数值变量为 1 至 22 的整数值，月份的调整方式与时间变量类似；第二，由于数据集是时间序列数据，本文将数据集的目标值设置为 0-1 变量，1 代表该供应商在下一期的销量是上升的，0 则代表该供应商在下一期的销量是持平或下降的。因为细分市场共有 4 类，所以用于模型训练的数据集共有 4 个。对于每个数据集来说，实验中聚合了观测时期内的所有数据，采用滑动窗口依据上文提到的 0-1 变量的定义来设置目标值 Y 。

5.2 模型选择

本研究的销量趋势预测模型基于极端梯度提升树（XGBoost 模型）建立。XGBoost 是梯度提升树（GBDT）的改进，是集成学习模型的一种，可以用来做回归和分类任务。XGBoost 算法在许多预测任务中都有良好的表现。许多历史研究也都基于该模型并取得的良好预测效果，例如 Rafieian 等人在 2020 年将 XGBoost 用于估计广告与印象的匹配值中^[51]。Zhang 等人在 2022 年使用 XGBoost 预测餐厅的生存状况^[52]。

本研究选用 XGBoost 实施预测任务主要原因如下：首先，XGBoost 在高维特征数据（这些数据很可能存在互相关联性）中表现良好，而在本研究的 42 个变量中很可能存在互相关联性；其次，XGBoost 是贪婪的模型，它会选择含有最重要信息成分的数据进行预测，这将有助于本文进一步探索相似度指标对于互补者销量的重要性程度；最后，XGBoost 可以灵活地处理预测因子之间存在潜在的高纬度相关性，例如，相似度指标的提升可能对长期专注于该细分市场的互补者具有更多的影响性。同时，实验中也尝试了其他预测模型（随机森林、SVM 等）的预测效果，结果如表 5.1 所示，结果表面 XGBoost 具有最优的预测性能。

表 5.1 XGBoost 与其他模型的预测结果对比 (AUC)

	XGBoost	Random forest	SVM
Baseline	0.7614	0.7346	0.7127
Baseline+similarity	0.7789	0.7491	0.7301

5.3 模型训练及参数调整

定义 i 为供应商个体, t 为时期, 本研究的 XGBoost 模型定义如公式 (7) 所示。该公式包含两部分: $loss(\theta)$ 和 $\Omega(\theta)$ 。前者是损失函数, θ 是需要调整的模型参数的集合, 后者为正则化项, 用来控制模型的过拟合。 $X_{it} = (x_{it-l}, x_i, x_{t-l})$ 是自变量的向量集, x_{it-l} 表示每个供应商与时间相关的变量 (例如评论数量、产品数量等); x_i 表示与供应商个体无关的时间相关变量 (例如季节, 该时期内供应商总数等); x_{t-l} 表示与供应商有关与时间无关变量 (例如产品类别)。

$$\begin{aligned}
 loss(\theta) &= - \sum_{it}^{\min_{\theta} loss(\theta) + \Omega(\theta)} [y_{it} \ln \hat{y}_{it} + (1 - y_{it}) \ln(1 - \hat{y}_{it})] \\
 \Omega(\theta) &= \sum_{g=1}^G [\gamma L_g + \frac{1}{2} \gamma \sum_{l=1}^{L_g} \omega_{gl}^2] \\
 \hat{y}_{it} &= \frac{e^{\sum_{g=1}^G f_g(X_{it-l})}}{1 + e^{\sum_{g=1}^G f_g(X_{it-l})}}; f_g : X_{it-l} \rightarrow \omega_{gl}, l = 1, 2, \dots, L_g; \\
 &g \text{ 为树, } l \text{ 为叶子}
 \end{aligned} \tag{5-1}$$

XGBoost 模型的调参步骤如下: (1) 根据表 3-2 中的参数设定范围, 首先使用网格搜索法确定树的最大深度 max_depth 和最小叶节点权重 min_child_weight ; (2) 基于 (1) 选定的值, 使用同样的方法确定 $colsample_bytree$ 和 $subsample$ 的值; (3) 调整 $gamma$ 参数; (4) 设定学习率 $learning_rate$ 。我们模型的最终调参结果如表 5.2 所示。

表 5.2 模型参数设定

参数	参数定义	预设值	终值
max_depth	树的最大深度	[3, 4, 5, 6]	4
min_child_weight	最小叶子节点中样本的权重和	[1, 2]	2
$colsample_bytree$	每棵树随机采样的列数占比	[0.6, 0.8, 1]	1
$subsample$	每棵树随机采样的比例	[0.6, 0.8, 1]	0.8
$scale_pos_weight$	控制正负样本权重的平衡, 通常用于不均衡分类	[1]	1
$learning_rate$	学习率, 控制模型的收敛步长	[0.05, 0.1, 0.3, 0.5]	0.2
$gamma$	树分裂时损失函数的最小下降值, 参数越大代表算法越保守	[0, 4, 8]	4

5.4 模型结果评估

根据本研究的细分市场类别，实验中共训练了 4 个模型；同时，实验基于排除互补者与平台产品相似度的数据训练了 4 个基线模型以作对比。为了保证模型的鲁棒性，每个模型均采用了十折交叉验证。模型的结果如表 5.3 所示，模型结果对比的 ROC 曲线如图 5.1 所示。

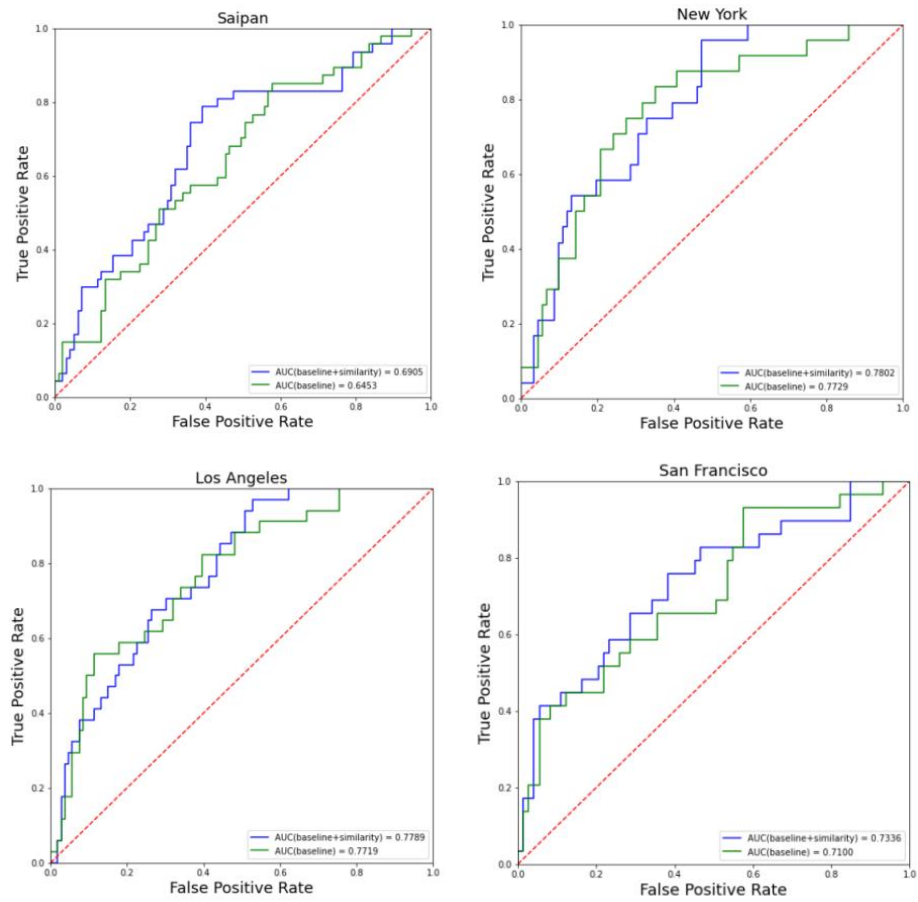


图 5.1 模型结果对比（ROC 曲线）

表 5.3 模型结果对比

	模型	评价指标					
		AUC	Accuracy	Precision	Recall	F1	KS
整体	Baseline	0.7155	0.7301	0.5501	0.3684	0.4395	0.2485
	Baseline+Similarity	0.7389	0.7638	0.6416	0.4493	0.5257	0.3404
洛杉矶	Baseline	0.7339	0.6911	0.5714	0.3373	0.4242	0.2085
	Baseline+Similarity	0.7344	0.7317	0.6441	0.4578	0.5352	0.3290
塞班岛	Baseline	0.6453	0.6944	0.5417	0.3404	0.4211	0.2064
	Baseline+Similarity	0.6905	0.6945	0.5415	0.4255	0.4762	0.2503
旧金山	Baseline	0.7100	0.7549	0.6111	0.3793	0.4681	0.2834
	Baseline+Similarity	0.7336	0.7941	0.7500	0.4138	0.5333	0.3590
纽约	Baseline	0.7729	0.7826	0.4762	0.4167	0.4444	0.2958
	Baseline+Similarity	0.7972	0.8348	0.6316	0.5000	0.5581	0.4231

5.5 稳健性检验

为了使模型在样本外数据得到更稳定的表现，本节使用了两种方法来检验模型的稳健性：

第一，减去重叠效应，替换累计变量的计算方式。具体操作为，将之前的 $t-1$ 的值和 $t-1$ 时期的累计均值改为了 $t-1$ 时期的值和 $t-2$ 时期的累计均值。

第二，更换互补者与供应商产品相似度指标的计算规则。具体操作为，更换 4.2 节中描述的互补者与供应商产品相似度的算法，将其改为基于词袋模型的独热向量的余弦相似度。模型稳健性检验结果如表 5.4 所示。

表 5.4 模型稳健性检验结果

稳健性检验方式	模型（整体）	评价指标					
		AUC	Accuracy	Precision	Recall	F1	KS
替换累计变量	Baseline	0.6649	0.7662	0.6000	0.3000	0.4000	0.2298
	Baseline+Similarity	0.6721	0.7549	0.6000	0.4138	0.4898	0.3042
替换产品相似度	Baseline	0.6535	0.7033	0.6087	0.3373	0.4341	0.2269
	Baseline+Similarity	0.6798	0.7410	0.5393	0.3529	0.4267	0.2397

第六章 基于 SHAP 的模型解释性分析

在机器学习领域，人们通过数据来训练模型以达成良好的预测和分类任务已经得心应手。近些年来，关于可解释性机器学习的研究也愈发普遍，不管是传统的机器学习模型还是深度学习模型，人们不再在精确度和可解释性之前寻求平衡，而是希望在获得高准确度的模型结果同时也了解模型是如何基于数据做出预测的。关于模型可解释性的方法有很多，目前基于 SHAP 的模型可解释分析方法被广泛使用。

本章将基于 SHAP 分析销量趋势预测模型的可解释性：（1）在 6.1 节中，总体概览了四个市场下每个模型 42 个特征中贡献度最高的 30 个特征的平均贡献程度，并得出对于变量重要性程度的初步分析；（2）6.2 节检验了高目的地城市相似度的数据对模型预测效果的影响，并对细分市场进行了更精细的划分；（3）6.3 节则在 6.2 节划分的细分市场基础上，基于 6.1 节的变量分析结果进行了相关性分析。其中，6.3.1 节探究了细分市场类别与产品相似度指标贡献程度的关系；6.3.2 节探究了不同产品规模的互补者对产品相似度指标的敏感程度。

6.1 特征重要性分析

SHAP 特征重要性（SHAP feature importance）本质上是一个预测器考虑到所有其他预测器的所有可能组合的边际贡献。本研究基于第五章建立的四个细分市场下的销量趋势预测模型，使用 python 的 shap 库，得出了四个细分市场下每个预测模型的 SHAP 特征重要性值，图 6.1 展示了 42 个变量中，贡献程度最高的前 30 个特征值的平均值。平均值计算的方法是首先计算出每个特征值在四个模型中的贡献百分比，然后将比值进行平均并与四个模型的所有的 SHAP 值相乘，继而求出每个特征的平均 SHAP 重要性值。

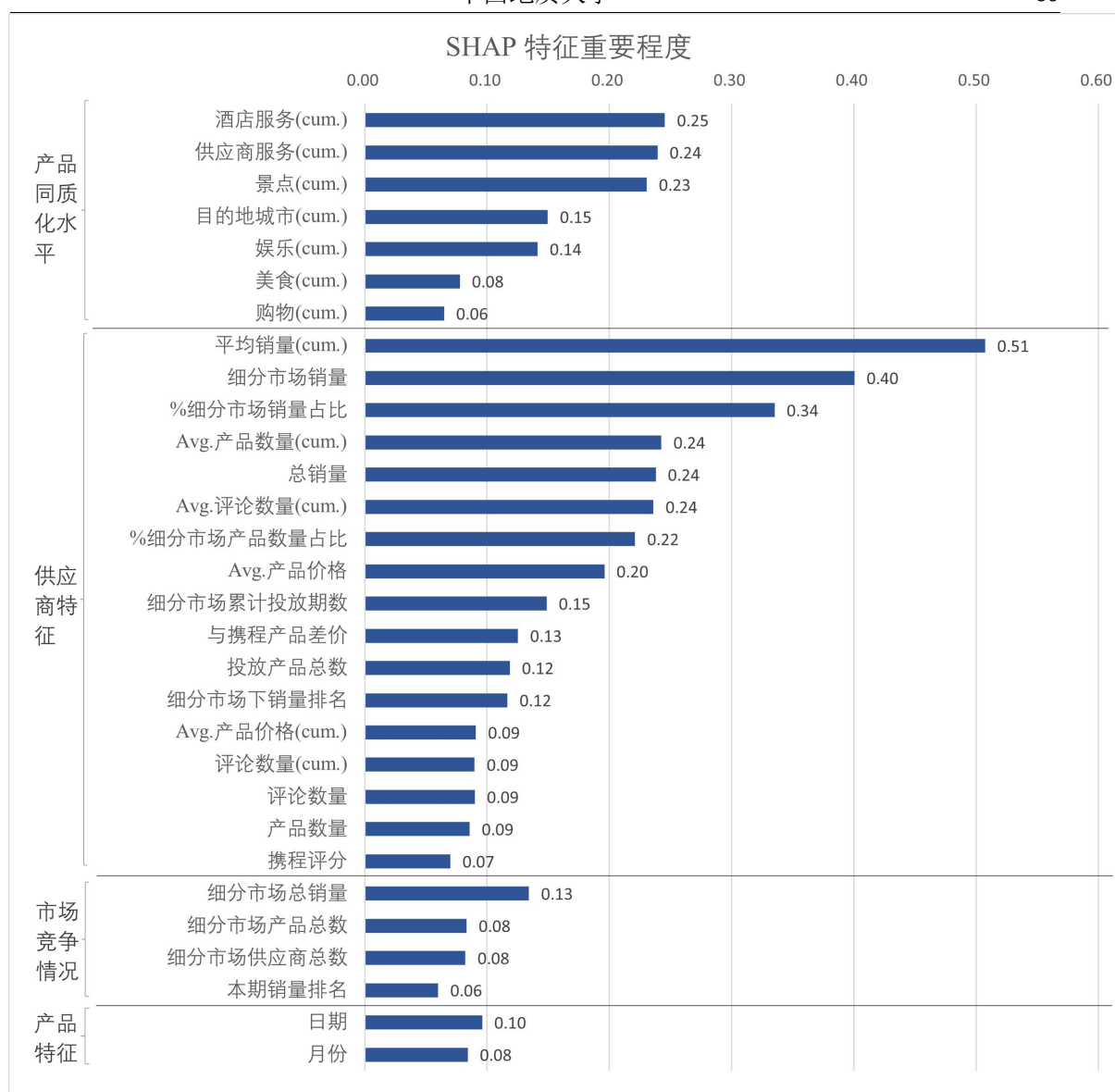


图 6.1 SHAP 特征重要性程度最高的前 30 个变量

注：本图中的 SHAP 值按照四个预测模型中每个变量的平均贡献比例汇总计算得出；出发城市、细分市场的变量作为控制变量固定

从结果来看，可以初步得到以下结论：（1）对模型最具贡献价值的是分别是平均销量、细分市场销量和细分市场销量占比这三个指标，且整体来说，累计平均值的贡献价值要大于单期的指标。（2）本研究重点关注的互补者与平台产品同质化水平的指标中，酒店服务、供应商服务和景点的相似度指标对模型的贡献也较大，超越了 70% 的特征。这进一步验证了 5.4 节中提及的模型评估结果：产品相似度指标对模型的预测结果具有重要的贡献。（3）对于互补者来说，与平台产品的相似度指标在很大程度上可以成为其预测自身未来销量走势的考虑因素。但需要注意的一点是，产品相似度指标中只有三个指标的贡献比较明显，本

研究统计了每个产品维度的内容在产品中的占比，娱乐、美食、购物、航空服务这四个维度的内容在整体产品中占比较小，计算出的相似度结果值更集中，模型无法很好的基于这些值做出优秀的预测，因此这些指标在模型中的贡献度较小。

6.2 目的地城市相似度与模型预测效果

在图 6.1 中，重要性程度仅次于酒店服务、供应商服务和景点相似度指标的产品相似度变量是目的地城市。考虑到 4.3.1 节定义产品特征变量时，本文仅基于携程网站上标注的目的地城市信息来划分细分市场，但在实际情况中，归属于同一目的地城市分类下的产品的实际目的地城市路线并不相同。因此，本研究猜想，是否产品相似度指标对于目的地城市相似度高的产品具有更高的预测价值。为了验证这个猜想，在此基于目的地城市相似度这一指标，依照其中位数为界划分为了两个子数据集：一个是高目的地城市相似度的数据集，另一个是低目的地城市相似度的数据集。在具体的实验操作中，本研究为洛杉矶、旧金山、纽约这三个细分市场分别拆分了两个子数据集，共得到 6 个子数据集进行模型训练。塞班岛的数据在这里没有被继续拆分是因为塞班岛的相关旅行产品的目的地城市几乎没有差别（由于其地理位置的原因）。

经过对比，研究发现对于高目的地城市相似度的子数据集，模型效果要远远大于相对于的低目的地城市相似度的数据集，实验结果在表 6.1 中展示。

表 6.1 不同目的地城市相似度数据的模型效果对比

子数据集类别 (目的地城市相似度水平)		评价指标					
		AUC	Accuracy	Precision	Recall	F1	KS
整体	高	0.7383	0.7899	0.7746	0.5847	0.6629	0.4766
	低	0.5238	0.7294	0.2595	0.1815	0.3107	0.3124
洛杉矶	高	0.7032	0.7183	0.6667	0.6207	0.6429	0.4064
	低	0.4806	0.7536	0.1000	0.1111	0.1053	0.3890
旧金山	高	0.7677	0.8039	0.8571	0.6000	0.7059	0.5355
	低	0.4833	0.6275	0.2500	0.1333	0.1739	0.3333
纽约	高	0.7439	0.8475	0.8000	0.5333	0.6400	0.4879
	低	0.6074	0.8070	0.4286	0.3000	0.3529	0.2149

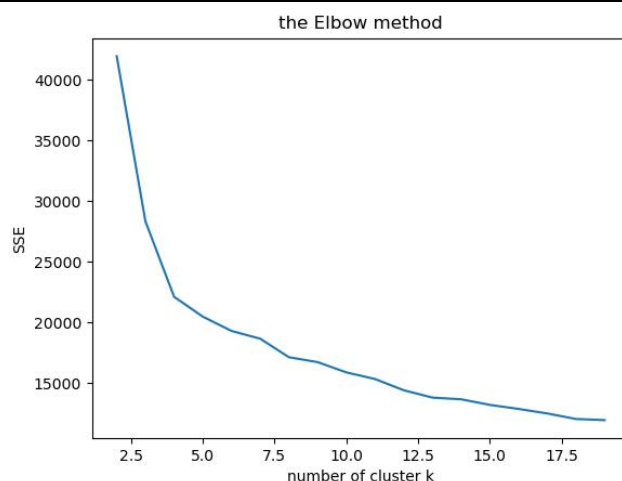


图 6.2 K-Means 无监督聚类学习模型聚类系数 k 与 SSE 关系

根据以上发现，本研究使用无监督学习将产品进行了更细层次的聚类，希望将高相似度的目的地城市路线的产品聚为一类，从而进行进一步探索。本研究使用了经典的无监督聚类模型 K-Means 对产品基于目的地城市进行聚类。首先，将所有的目的地城市放入一个词袋中，然后使用独热编码（One-Hot）将每个产品表示为一个数值向量，向量的维数为 127（代表了 127 个目的地城市）。接着，本研究根据 K-Means 模型聚类系数 k 的选择方法肘部原则（如图 6.2）决定将聚类系数 k 定为 12，即将细分市场的类数定为 12 类，然后基于现实意义人工将类别进行合并，总共得到了 7 类细分市场，重新定义的细分市场的如表 6.2 所示。

本文后期在进行实验的过程中，选取了产品数量占比最多的前三类市场进行分析，其分别是：洛杉矶、旧金山；波士顿、纽约；塞班岛。这表示了在之前的细分市场分类的基础上，通过聚类发现洛杉矶和旧金山这两个目的地城市更多地出现在同一产品中，例如纽约和波士顿通常伴随出现。此步骤筛选出进行进一步分析的这三类产品占据所有产品总量的 58%。

表 6.2 人工合并后的旅游产品细分市场类别

细分市场编号	目的地城市	产品数量	产品占比
0	洛杉矶、旧金山	11170	26%
1	波士顿、纽约	7966	19%
2	塞班岛	5696	13%
3	迈阿密、奥兰多	2013	5%
4	西雅图、波兰特	1336	3%
5	夏威夷	1176	3%
6	关岛	703	2%

注：聚类之后会存在一部分噪声数据，本研究统计了聚类结果总数较多的细分市场类别

6.3 特征相关性分析

SHAP 值处理衡量重要性之外，还可以得出特征与预测结果的相关性。6.1 节对预测模型的特征贡献度结果做了基础的评估，为了研究与平台产品相似度对互补者销量趋势的影响，本节将会地理位置的原因几乎所有的产品都只会单独包含塞班岛这个目的地城市。通过重新聚类进一步探究关键特征值间的相关关系。在 42 个变量中前 15 个对销量趋势预测最有价值的变量中，可以看出，这些变量除了传统研究中已知的对销量预测的有利因素（销量、价格、评论数）和本实验假设的产品相似度指标外，还有一些值得进一步研究的变量对预测的帮助性较大，本节将对这些不能直观解释的变量做进一步分析。

6.3.1 细分市场类别与产品同质化水平

尽管 6.1 节得到了产品相似度指标的特征重要性程度，实际上，对于每个不同的细分市场，其产品相似度特征的 SHAP 贡献程度都是不同的。本研究观察到，对于纽约、波士顿这个细分市场来说，景点相似度的贡献值大于供应商服务，供应商服务相似度大于酒店服务相似度。而对与另外两个细分市场，相似度指标的贡献程度又是不同的。

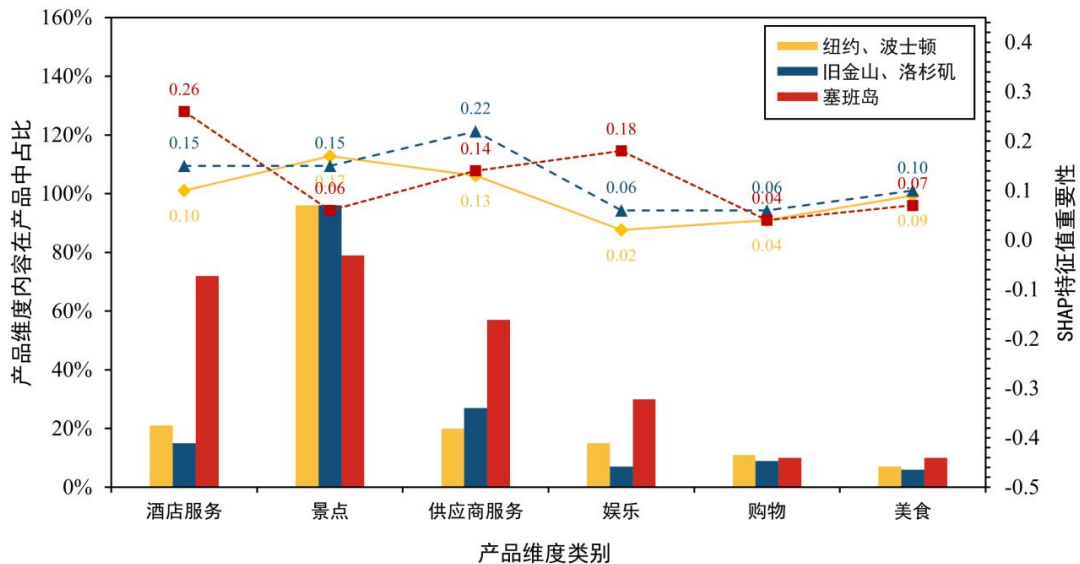


图 6.3 不同细分市场下产品维度内容占比与 SHAP 特征值重要性程度的关系

如图 6.3 所示，对旧金山、波士顿这个细分市场来说，供应商服务对其贡献价值最高，而对于塞班岛来说，酒店服务的贡献价值最高。为了探究这种不同产

品维度类别的贡献程度的区别，本研究继续统计了每个细分市场下每个产品维度内容在产品中的占比。从图 6.3 中可以看出，当产品维度内容在产品中占比越高时，该产品维度将会倾向于有更高的 SHAP 特征值。在实际情况里，对于纽约、波士顿和旧金山、洛杉矶这两个细分市场来说，其景点种类繁多，因此景点相似度对销量趋势预测的贡献更大。而塞班岛细分市场的景点内容相对更少，其产品更多侧重于娱乐和酒店服务的内容，因此这两者对于销量趋势预测的贡献更大。总体来说，产品维度类别对旅游产品销量趋势预测的贡献程度并不是恒定的，和具体细分市场的产品内容特点有关。

6.3.2 互补者规模与产品同质化水平

本节探究了互补者规模与产品同质化水平对销量趋势预测模型的重要性程度的关系。这里使用了“产品数量”这一特征变量来作为衡量互补者规模的因素，原因如下：（1）一般情况下我们认为，规模越大的互补者会越倾向于投放更多的产品在携程平台；（2）产品数量在图 6.1 中的特征重要性较高，达到了第四名，因此使用该特征值会得到更好的分析效果。为了对比产品相似度相关的特征值与传统销量趋势预测指标的重要性程度，本研究选择了除了目的地城市相似度之外的最高的四个产品相似度特征作为产品相似度指标的衡量特征，分别是酒店服务、供应商服务、景点、娱乐、购物、美食维度（这六个指标均选取了累计平均值而非单期值）。另外，本研究选取了过去平均评论数、过去平均销量、与携程产品差价来表示传统的销量趋势预测特征。除此之外，研究还加入了一个市场竞争的特征，即细分市场总销量，因为该特征值在图 6.1 中显示的重要性程度位居第二。

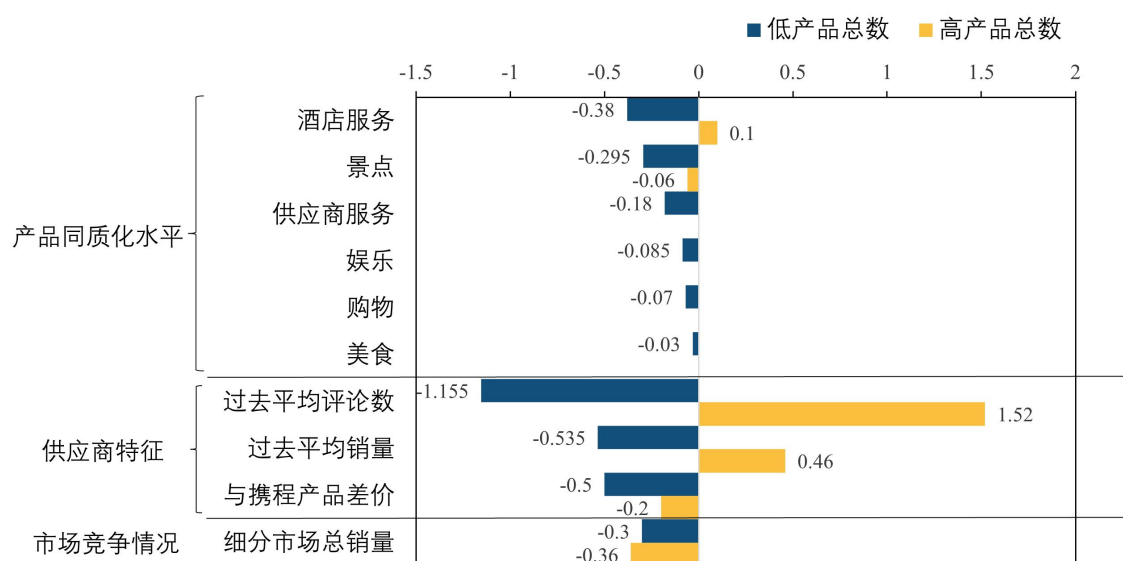


图 6.4 产品规模与变量的 SHAP 值

SHAP 值的优越之处在于除了可以给出特征的重要性程度之外，还可以得出特征与预测结果的正向、负向相关性。因此，基于 6.2 划分的三个细分市场，本研究按照产品总数的中位数将数据集划分为了两个子数据集，分别是低产品总数的供应商数据和高产品总数的供应商数据。最后，共得到 6 个子数据集，本文针对每个子数据集进行了 XGBoost 模型训练并得到我们观测指标的 SHAP 值，并将 SHAP 值进行平均得到图 6.4 所示的数据。

根据图 6.4 可以看出，对于高产品总数的互补者来说（即产品规模较大的互补者），其销量趋势预测更依赖于传统指标，即评论数、销量、价格等因素，这些指标对于高产品总数的互补者均是正向促进作用，即这些指标的数值越大，互补者未来的销量趋势更可能呈现上升趋势；对于低产品总数的供应商来说，其销量趋势预测除了基于传统指标之外，产品相似度指标也贡献了一定的预测价值，并且产品相似度指标与销量趋势的预测呈负相关，这在一定程度上表面对于产品规模较小的互补者来说，产品相似度越低将越有可能导致其未来销量趋势的上升。另外，值得注意的是对于过去平均评论数和过去平均销量这两个指标来说，其对于低产品规模互补者的未来销量趋势呈负相关关系，这是因为本研究中所使用的模型的目标值表示的是未来销量的升高或降低，对于低产品规模互补者来说，其销量相对于高产品规模的互补者更低，所以未来销量上升的可能性较大。总体来说，本节的探究进一步说明携程平台的产品同质化指标可能对于产品规模较小的互补者更具预测价值。

第七章 结论与展望

7.1 主要研究结论

在旅游电商平台的竞争环境下，本研究从细粒度的产品文本层面衡量了互补者与平台之间的产品同质化程度，探索了互补者与平台产品的产品同质化程度是否可以成为其预测销量趋势的指标这一问题。本文通过对旅游产品大规模文本数据的细粒度维度划分以及文本信息抽取，从产品特征的微观角度体现了供应商之间的产品同质化水平，从而构建 XGBoost 销量趋势预测模型，利用 SHAP 进行模型特征的可解释性分析，探究了与平台产品的同质化水平是否对互补者未来销量趋势预测有重要作用。通过实验，本研究的主要结论如下：

第一，与平台产品的同质化水平的相关指标在互补者未来销量趋势预测中起到了重要的作用，超过了 70% 的变量；这意味在平台强主导性的电商平台环境中，互补者在对自己未来销量趋势进行预测的过程里，除了可以将传统变量（如评论数、过去销量、产品价格等）考虑在内，还可以考虑与平台产品的同质化水平。

第二，在互补者与平台产品同质化水平的相关指标中，不同的旅游细分市场中不同维度的指标对于预测模型的贡献程度是不同的，主要受到该细分市场不同的产品维度内容占比的影响；因此，互补者应对旅游市场中不同细分市场的差别保持敏感性，考虑细分市场差别从而对销量趋势进行更精准的判断。

第三，不同产品规模的互补者在销量趋势预测中对与平台产品同质化水平的相关指标的依赖性是不同的，产品规模较大的互补者更依赖传统的销量趋势预测指标（评论、价格、历史销量水平等），产品规模较小的互补者除了依赖传统指标之外，与平台产品的同质化水平相关的指标也对预测模型有着较大的贡献，并且这些产品同质化水平的指标与互补者销量呈现负相关关系。因此，对于产品规模较小的互补者来说，其未来销量走势可能和产品同质化水平指标的相关性更大。

7.2 管理建议

本研究围绕电商平台中互补者的产品特征对其销量趋势预测的作用这一问题展开，关注了平台强主导性的影响下，探究了互补者与平台产品的同质化水平是否可以作为互补者自身预测未来销量趋势的关键指标这一问题。根据研究结论，本文提出了以下管理建议以供参考：

第一，对于互补者来说，电商平台中存在着高透明、多方参与的复杂竞争。在互补者考虑入驻电商平台之前，首先可以根据自身特点、参考平台产品与自身的同质化特征，决定是否进驻平台以及预估在平台的生存状况。另外，互补者在激烈的市场竞争中需要根据自身的规模情况等制定恰当的产品策略，以此在市场中最合适的自身定位。事实证明，消费者在进行消费选择时，更可能优先选择知名企业的产品。因此，尤其对于产品规模较小的互补者需要尽可能打造差异化竞争，避免与平台所有者产品同质化水平过于相似而导致自身处于竞争劣势。

第二，对于平台所有者来说，虽然其产品在平台市场中占有强大的优势，但仍要考虑到整个平台的生态治理问题，因为平台的活性既取决于吸引消费者的能力，也取决于吸引互补者入驻的能力。因此，平台所有者即使具有优厚的资源以支持其产品的快速更新和迭代，但也需要考虑到整个平台生态，尽量避免占据过多的互补者市场继而抑制其创新活性甚至导致互补者退出平台等。

7.3 不足与展望

本研究中存在的不足与相关展望如下：

第一，扩大样本数据的时间窗至以“年”为单位的周期。由于本研究所使用的实验数据是以“月”为时间周期，且时间跨度仅两年，因此数据特征的平稳性略欠缺，无法捕捉数据的长期变化规律。因此，未来的研究可以使用更长时间跨度的数据，例如以“年”为单位的时间窗，从而获得更加稳健的模型结果。

第二，本研究未考虑宏观政治政策、行业政策、平台或企业自身政策等因素对互补者销量趋势的影响。未来的研究可以加入这些变量进行综合考虑，以期得到更完整和严谨的研究结论。

第三，与平台所有者的产品同质化水平对互补者销量趋势预测的影响关系有待进一步细化研究。本研究基于 SHAP 对预测模型进行了可解释性分析，探究了细分市场类别与互补者规模的差异对产品同质化指标的预测价值的相关性。未来可以使用因果识别等方法，进一步探究相关性的具体程度以及因果效应。

第四，本研究结论在其它领域价值可用性未知。本研究选取了 2017 年 3 月至 2018 年 12 月的美国出境游旅行产品文本数据进行分析与销量趋势预测模型的构建。结果能够为互补者在平台中的产品投放、销量趋势预测以及平台的生态治理等提供一定建议和管理价值，但该研究结论是否适用于其它旅游市场以及电商平台，仍待考证。未来可以进一步尝试对比更多平台的数据及结果，从而完善结论的全面性。

致谢

本科四年终场，首先想对自己说一声谢谢。

感谢自己始终充满希望，感谢自己蓬勃的生命，感谢自己永远相信自己，也感谢自己每一次勇敢的抉择与坚持的勇气。

其次，也感谢四年来所有给予我帮助和鼓励的人。感谢我的父母和我的哥哥嫂子，永远爱我也坚定地支持我的每一个决定！感谢朱镇老师带领大二的我进入科研的大门，耐心地指导我读文章、做研究；也很感谢石咏、刘保山、王飞老师在学业方面给予我沟通、帮助和指点；也感谢四年来每一位任课教师对我的教导！感谢我的叔叔徐荣锴从不厌倦地和我交流以及分享知识；感谢数据科学家 Tuso 老师对我在实习过程中的帮助；感谢我在普华永道的实习导师 Micheal Liu 在工作中对我的指引，感谢您对我的欣赏和鼓励；感谢我的闺蜜胡秋雨在大学听我讲了四年的废话；大学四年里结交了很多校内校外的朋友，非常感谢和朋友们的沟通与交流，让我拥有认知的拓展和思想的碰撞；非常感谢其他这四年里所以帮助过我的人，因为你们，才有了今天更好的我！

回想四年来无数日夜，恍如天上虚无飘渺的月，仿佛近在眼前，又似乎越来越远。顺着时光的长河慢溯，我又看见父母送我第一次来武汉时满城灿烂的月季，它们见证了我 17 岁那年幼稚又渴求的心。在南望山淋漓的秋雨和落叶弥漫的天空里，我体验了第一次进入大学的新奇。我站在长江边上，轻柔的晚风陪着我散步——那一刻我真希望可以永远留住这无尽的美好。可时间的长河却开始了激流勇进，我跑向前去：每一个站在图书馆门口望向天空的夜晚被我尽收眼底；那些孤独的日子、每个摸索前行的黑夜里，我的心就是一盏灯，忽明忽暗却永远亮着。我看见了每一次的尝试、成功和失败，也好像重新感受了每一次的渴望、开心和叹气。还有那一路上许许多多的人啊，我们相遇或擦肩而过。后来，我站在上海的外滩边上，高楼林立间，我第一次觉得自己渺小；就像在北京的寒风里，我的周围人潮汹涌。而我的偏执似乎是一个疯狂的巨人，我挤进人群又穿过人群，却发现抬头是很大一片天空。说着说着，时光的船就这样驶到了眼前，也许这就是我生命中最美好的时光了吧，我缓过神来望向前去，只发觉此行山高路远，请一定不要辜负自己。

参考文献

- [1] Jacobides M G, Cennamo C, Gawer A. Towards a theory of ecosystems[J]. Strategic management journal, 2018, 39(8): 2255-2276.
- [2] Nambisan S, Siegel D, Kenney M. On open innovation, platforms, and entrepreneurship[J]. Strategic Entrepreneurship Journal, 2018, 12(3): 354-368.
- [3] Rochet J C, Tirole J. Platform competition in two-sided markets[J]. Journal of the european economic association, 2003, 1(4): 990-1029.
- [4] Nalebuff B J, Brandenburger A M. Co-opetition: Competitive and cooperative business strategies for the digital economy[J]. Strategy & leadership, 1997, 25(6): 28-33.
- [5] Wang Q, Li B, Singh P V. Copycats vs. original mobile apps: A machine learning copycat-detection method and empirical analysis[J]. Information Systems Research, 2018, 29(2): 273-291.
- [6] Huang H J, Yang J, Zheng B. Demand effects of product similarity network in e-commerce platform[J]. Electronic Commerce Research, 2021, 21: 297-327.
- [7] Smith S L J. The tourism product[J]. Annals of tourism research, 1994, 21(3): 582-595.
- [8] Xu J B. Perceptions of tourism products[J]. Tourism management, 2010, 31(5): 607-610.
- [9] Moore J F. Predators and prey: a new ecology of competition[J]. Harvard Business Review, 1993, 71(3): 75.
- [10] 陈威如,余卓轩.平台战略：正在席卷全球的商业模式革命[M].北京:中信出版社,2013:268.
- [11] Iansiti M, Levien R. Strategy as ecology[J]. Harvard business review, 2004, 82(3): 68-78.
- [12] Tiwana A. Platform ecosystems: Aligning architecture, governance, and strategy[M]. Newnes, 2013.
- [13] Athey S, Roberts J. Organizational Design: Decision Rights and Incentive Contracts[J]. American Economic Review, 2001, 91(2): 200-205.
- [14] Eisenmann T, Parker G, Alstyne M V. Platform Envelopment[J]. Strategic Management Journal. 2011, 32(12): 1270-1285.
- [15] Gawer A, Cusumano M A. Industry platforms and ecosystem innovation[J]. Journal of product innovation management, 2014, 31(3): 417-433.
- [16] Cenamor J. Complementor competitive advantage: A framework for strategic decisions[J]. Journal of Business Research, 2021, 122: 335-343.
- [17] Li Z, Agarwal A. Platform integration and demand spillovers in complementary markets: Evidence from Facebook's integration of Instagram[J]. Management Science, 2017, 63(10): 3438-3458.

- [18] Jacobides, M. G. & Cennamo, C. & Gawer, A. Towards a Theory of Ecosystems [J]. *Strategic Management Journal*, 2018, 22 (08) : 2255-2276.
- [19] Cennamo, C. Building the Value of Next-generation Platforms: the Paradox of Diminishing Returns [J]. *Journal of Management*, 2018, 44 (08) : 3038-3069.
- [20] Park C W, Milberg S, Lawson R. Evaluation of brand extensions: The role of product feature similarity and brand concept consistency[J]. *Journal of consumer research*, 1991, 18(2): 185-193.
- [21] Muthukrishnan, A. V., & Weitz, B. A. (1991). Role of product knowledge in evaluation of brand extension. *Advances in Consumer Research*, 18(1), 407-413.
- [22] Bangyong, L., Juanzi, L. I., & Kehong, W. (2004). Web page recommendation model for the semantic web. *Journal of Tsinghua University*, 44(9), 1271-1272.
- [23] Zhai Z, Liu B, Xu H, et al. Clustering product features for opinion mining[C]//*Proceedings of the fourth ACM international conference on Web search and data mining*. 2011: 347-354.
- [24] Chang C C, Lin B C, Chang S S. The relative advantages of benefit overlap versus category similarity in brand extension evaluation: The moderating role of self-regulatory focus[J]. *Marketing Letters*, 2011, 22: 391-404.
- [25] 夏季.平台互补者的产品差异化定位策略与绩效的关系研究[D].浙江工商大学, 2021.DOI:10.27462/d.cnki.ghzhc.2021.000830.
- [26] Colucci L, Doshi P, Lee K L, et al. Evaluating item-item similarity algorithms for movies[C]//*Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 2016: 2141-2147.
- [27] Zhou W, Mok P Y, Zhou Y, et al. Fashion recommendations through cross-media information retrieval[J]. *Journal of Visual Communication and Image Representation*, 2019, 61: 112-120.
- [28] Grishman R . Message Understanding Conference-6: A Brief History[C]//*Proceedings of the 16th conference on Computational linguistics*.1996.
- [29] Chinchor N . MUC-7 Named entity task definition version 3.5[J].1997.
- [30] Xie R, Liu Z, Jia J, et al. Representation learning of knowledge graphs with entity descriptions[C]//*Proceedings of the AAAI Conference on Artificial Intelligence*. 2016, 30(1).
- [31] Babych B, Hartley A. Improving machine translation quality with automatic named entity recognition[C]//*Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*. 2003.
- [32] Zhu J, Uren V, Motta E . ESpotter: Adaptive Named Entity Recognition for Web Browsing[C]// *Third Biennial Conference on Professional Knowledge Management*. 2005.
- [33] Lample G, ballesteros M, subramanian S, et al, Neural Architectures for Named Entity Recognition[J].*arXiv preprint arXiv : 1603.01360*,2016.
- [34] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named

- entity recognition[J]. arXiv preprint arXiv:1603.01360, 2016.
- [35] Žukov-Gregorič A, Bachrach Y, Coope S. Named entity recognition with parallel recurrent neural networks[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2018: 69-74.
- [36] Zhou J T,Zhang H, Jin D,et al. Roseq : Robust Sequence Labeling [J].IEEE Transactions on Neural Networks and Learning Systems,2019,PP (99) : 1-11.
- [37] Collobert R, Weston J,Bottou L,et al. Natural Language Processing (almost) from Scratch[J].Journal of Machine Learning Research,2011,12 (Aug) : 2493-2537.
- [38] Ma X,Hovy E.End-to-end Sequence Labeling Via Bidirectional Lstm-cnns-crf[J].arXiv preprint arXiv :1603.01354,2016.
- [39] Liu L, Shang J,REN X,et al. Empower Sequence Labeling with Task-aware Neural Language Model [C] // Thirty-Second AAAI Conference on Artificial Intelligence.2018.
- [40] 王健宗, 孔令炜, 黄章成, 等.图神经网络综述 [J].计算机工程, 2021, 47 (4):1-12.
- [41] Chen T, Benesty M, He T (2018) Understand Your Dataset with Xgboost (R Document).
- [42] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[J]. Advances in neural information processing systems, 2017, 30.
- [43] Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J]. PloS one, 2015, 10(7): e0130140.
- [44] Datta A, Sen S, Zick Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems[C]//2016 IEEE symposium on security and privacy (SP). IEEE, 2016: 598-617.
- [45] Lipovetsky S, Conklin M. Analysis of regression in game theory approach[J]. Applied Stochastic Models in Business and Industry, 2001, 17(4): 319-330.
- [46] Ribeiro M T, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016: 1135-1144.
- [47] Shrikumar A, Greenside P, Shcherbina A, et al. Not just a black box: Learning important features through propagating activation differences[J]. arXiv preprint arXiv:1605.01713, 2016.
- [48] Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions[J]. Knowledge and information systems, 2014, 41: 647-665.
- [49] Molnar C. Interpretable machine learning[M]. Lulu. com, 2020.
- [50] Chevalier J A, Mayzlin D. The effect of word of mouth on sales: Online book reviews[J]. Journal of marketing research, 2006, 43(3): 345-354.
- [51] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data

-
- mining. 2016: 785-794.
- [52] Rafieian O, Yoganarasimhan H. Targeting and privacy in mobile advertising[J]. Marketing Science, 2021, 40(2): 193-218.
- [53] Zhang M, Luo L. Can consumer-posted photos serve as a leading indicator of restaurant survival? Evidence from Yelp[J]. Management Science, 2023, 69(1): 25-50.

附录 A：命名实体识别过程的技术细节

附录 A 将介绍本研究中基于深度学习模型 BERT-LSTM-CRF 对旅游产品文本数据进行命名实体识别的具体技术细节：

（1）文本标注过程

本研究使用 BIO 模式来标记产品数据中的实体。对于每个实体，第一个单词标记为“B-(实体名称)”，中间和末尾的单词标记为“I-(实体名称)”，非实体标记为“O”。BIO 的标注策略如表 A.1 所示。

实验中从整个数据集中随机选择了 2000 条数据作为训练集进行标注。本研究为了尽可能避免主观因素对数据标注过程的影响，标注结束后继续对标注人员的标注结果进行了抽样。随机抽取 1000 条标注数据，评估标注后的准确性。结果表明标注结果的正确率在 90%以上。

表 A.1 BIO 标注策略

子维度类别	Start Tag	Middle Tag	End Tag
Country	B-CNT	I-CNT	I-CNT
City	B-CY	I-CY	I-CY
Food	B-FOOD	I-FOOD	I-FOOD
Hotel_service	B-HOT	I-HOT	I-HOT
Airline_service	B-AIR	I-AIR	I-AIR
Attraction	B-ATT	I-ATT	I-ATT
Shopping	B-SHOP	I-SHOP	I-SHOP
Entertainment	B-ENT	I-ENT	I-ENT
Provider_service	B-PRO	I-PRO	I-PRO
Freedom_choice	B-FRE	I-FRE	I-FRE
Non entity tag	O	O	O

（2）模型训练及参数设置

许多复杂和深入的任务中经常引入新的命名实体，这需要重新训练 NER 模型来提取这些新实体。本研究基于前文介绍的标注策略使用标注好的训练数据训练了新的模型，这些实验都是在 NVIDIA Tesla K80 服务器上进行的，编程语言是 Python 3.7，使用的深度学习框架是 Tensorflow 1.13.1（这是一个由 Google 开发的深度学习框架）。本研究的深度学习模型参数设置如表 A.2 所示。

表 A.2 模型参数设置

参数名称	参数值
Optimizer	adam
Learning rate	0.001
Epoch	100
Dropout	0.5
Maximum-length sequence	300
Dimension of character embedding	100
Batch_size	128

(3) 模型评估结果

为了评估分类模型的性能，实验中将原始数据分为 75% 的训练数据和 25% 的测试数据。由于购物、美食维度的实体数量较少，这两个维度的评估效果略有下降，但整体来说模型的抽取效果较为准确。模型的整体评估结果如 A.3 所示。

表 A.3 模型评估结果

子维度类别	Precision/%	Recall/%	F1-Score/%
Destination-country	98.63	99.40	99.31
Destination-city	95.86	96.51	96.19
Attraction	88.07	88.54	88.30
Food	70.73	69.05	69.88
Entertainment	82.14	83.33	82.73
Shopping	62.07	73.47	67.29
Provider_service	76.72	86.83	81.46
Airline_service	92.38	97.00	94.63
Hotel_service	72.12	75.32	73.68
Freedom choice	93.31	91.20	91.20

(4) 实体合并结果

抽取出的实体中很可能存在着许多语义相似的实体。例如，“四星级酒店”、“四星酒店”、“四星级别酒店”这三个实体实际表达出了相同的意思。为了减少实体的数量，在构建图网络时使得产品实体之间尽可能连接相同的节点，实验中基于语义相似度对实体进行了进一步合并。具体的操作方法为，将每个实体由文本转化为数值形式上的 one-hot 向量，再使用余弦相似度计算实体间的相似度，大于一定阈值的实体将会被合并。实体合并前后的数量及合并时依据的语义相似度阈值设置如下表所示：

表 A.4 实体合并前后的数量对比及合并的语义相似度阈值

子维度类别	合并前实体数(个)	合并后实体数(个)	语义相似度阈值
Destination-country	9	4	0.7
Destination-city	233	127	0.7
Attraction	430	369	0.7
Food	39	33	0.7
Entertainment	98	62	0.7
Shopping	24	10	0.7
Provider_service	149	78	0.7
Airline_service	29	12	0.7
Hotel_service	102	62	0.7
Freedom choice	19	11	0.7

附录 B：基于 Neo4j 的供应商-产品网络可视化结果

附录 B 展示了 4.2 节中构建的供应商-产品网络图的实际可视化结果，图网络基于 Neo4j 图数据库构建，这里以旧金山的数据为例展示。

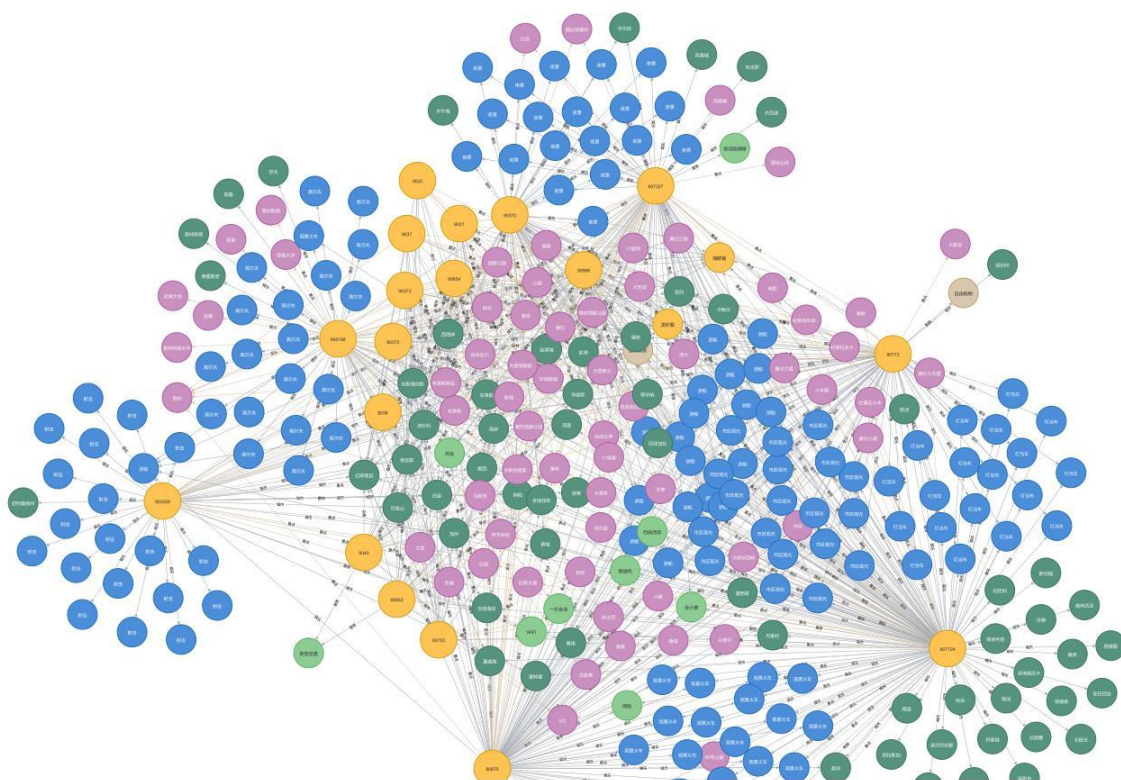


图 B-1 旧金山细分市场的供应商-产品网络 Neo4j 可视化展示

附录 C：变量统计汇总数据

附录 C 记录了用于训练 XGBoost 模型的变量统计汇总数据：

表 C.1 产品相似度指标统计信息汇总(累计值)

变量	计数	平均值	标准差	最大值	最小值
<i>Destination-city</i>	1665	0.68	0.22	0	1
<i>Attraction</i>	1665	0.67	0.29	0	1
<i>Hotel_service</i>	1665	0.73	0.36	0	1
<i>Airline_service</i>	1665	0.83	0.29	0	1
<i>Provider_service</i>	1665	0.62	0.38	0	1
<i>Shopping</i>	1665	0.84	0.31	0	1
<i>Entertainment</i>	1665	0.82	0.29	0	1
<i>Food</i>	1665	0.85	0.29	0	1
<i>Freedom_choice</i>	1665	0.80	0.25	0	1

表 C.2 供应商特征指标统计信息汇总

变量	计数	平均值	标准差	最大值	最小值
<i>number_of_periods</i>	1665	8.43	5.31	22	1
<i>submarket_product_num</i>	1665	8.58	16.39	177	1
<i>submarket_product_withsales_num</i>	1665	5.26	11.20	119	0
<i>total_product_num</i>	1665	86.32	177.42	1003	1
<i>total_product_withsales_num</i>	1665	60.12	122.29	713	0
<i>product_num_percentage</i>	1665	0.37	0.32	0	1
<i>price_avg</i>	1665	11347.04	9156.32	150000	1238
<i>price_avg_cum</i>	1665	11010.53	8647.95	142500	1345
<i>price_gap</i>	1665	6323.76	1460.01	11056	3052.47
<i>submarket_sales</i>	1665	27.90	176.16	5976	0
<i>submarket_sales_cum</i>	1665	31.57	67.05	930.71	0
<i>total_sales</i>	1665	295.91	1121.78	11914	0
<i>total_sales_cum</i>	1665	299.62	1002.16	10849.20	0
<i>product_sales_percentage</i>	1665	0.37	0.32	1	0
<i>comments_num</i>	1665	2.51	16.18	563	0
<i>comment_num_cum</i>	1665	3.08	6.38	98.5	1
<i>product_score</i>	1665	3.87	1.64	5	0
<i>product_score_cum</i>	1665	3.82	1.30	5	0

表 C.3 市场竞争情况指标统计信息汇总

变量	计数	平均值	标准差	最大值	最小值
<i>total_providers</i>	1665	26.18	10.81	49	3
<i>total_providers_haveSales</i>	1665	19.89	9.60	49	3
<i>total_products</i>	1665	186.75	118.78	524	6
<i>total_products_haveSales</i>	1665	159.89	101.13	524	6
<i>total_travle_count</i>	1665	954.58	1215.01	7233	0
<i>sales_rank</i>	1665	7.13	4.26	19	1