论文答辩

中国地质大学經濟管理學院
SCHOOL OF ECONOMICS AND MANAGEMENT, CHINA UNIVERSITY OF GEOSCIENCES

电商平台中互补者的产品同质特征与销量趋势预测研究

答辩人：徐嘉艺
指导教师：朱镇
2023年6月3日

目录
CONTENTS

中国地质大学經濟管理學院
SCHOOL OF ECONOMICS AND MANAGEMENT, CHINA UNIVERSITY OF GEOSCIENCES

1 研究背景与问题

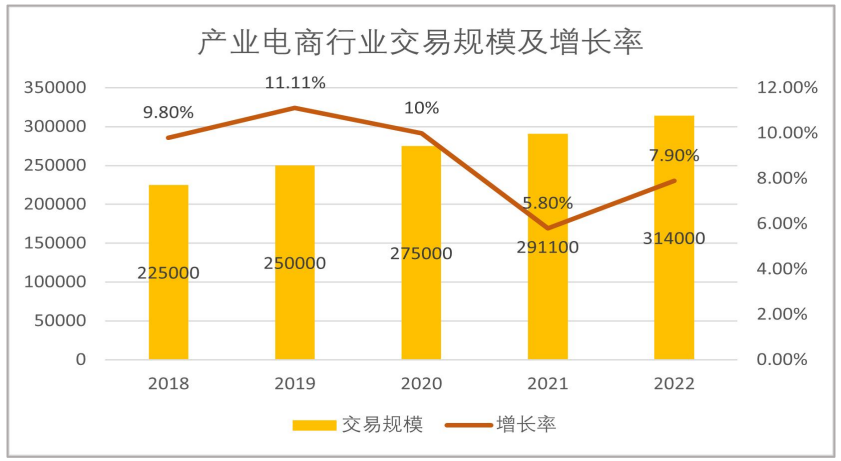
●实际背景 ●理论背景 ●研究问题

1.1 实际背景

电商大规模、高透明化的背景存在着**复杂激烈**的竞争，导致了**产品模仿**现象频繁出现

市场规模庞大、增势良好

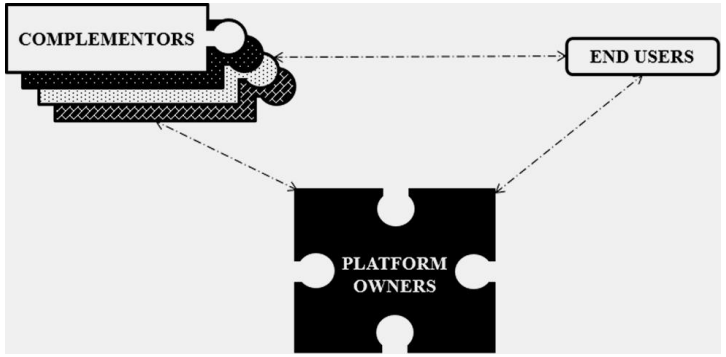
2022年中国电商市场规模31.4万亿元，同比增长7.86%[1]，虽受大环境影响行业整体增长预期不及去年，但总体增势依然强韧。因此，平台规模远超传统市场，吸引了更多的商家和用户。



产业电商行业交易规模及增长率[1]

互联网发展带来的高透明化背景下，平台存在更复杂的多方竞争关系

电商平台中更加透明化的竞争背景



平台生态中的三角相互依赖结构[2]

平台所有者：建设、维持、治理平台生态系统的企业。（亚马逊）
互补者：为平台提供互补品而为平台创造价值的企业。（亚马逊购物平台上除了亚马逊之外，同样进行商品销售的其他企业）

企业面临着严峻的挑战：如何在市场中找到自身产品定位？如何设计产品以获得更好的绩效表现？……

1 数据来自《2022年度中国产业电商市场数据报告》
2 Cenamor J. Complementor competitive advantage: A framework for strategic decisions[J]. Journal of Business Research, 2021, 122: 335-343.

1.2 理论背景

（1）从互补者视角出发的研究较少

关于平台的研究大部分从平台所有者的视角出发，重点关注了平台的成长与竞争等问题。有学者指出，互补者的行为及决策等研究是重要的研究方向[1]。

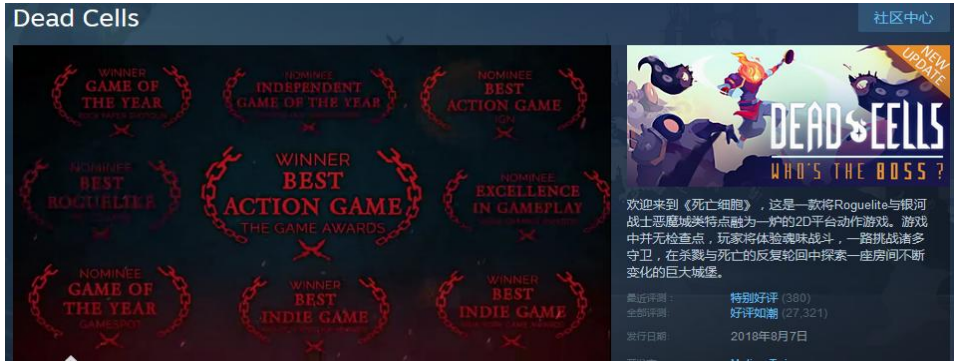
（2）很少有研究关注产品本身特征

有关**网络口碑**（在线文本评论[2]、图片评论[3]、社群行为[4]）等因素对企业绩效表现的研究已并不少见，但很少有研究关注产品本身，深入挖掘产品的特征及所蕴含的丰富价值。

（3）多数有关产品特征的研究缺乏细粒度探讨

尽管有学者从产品角度（产品相似网络[5]、文本描述相似度[6]等）探究了电商平台中企业的行为决策特征与其绩效的关系，但大多数研究仅及于TF-IDF等方法对产品文本进行了**数值意义上的相似度探究**，很少有研究将产品作为现实世界的物质实体来研究其更为丰富的细粒度信息。

某平台的在线文本评论与图片评论



steam页面上的游戏产品介绍

1 Cennamo C. Competing in digital markets: a platform-based perspective[J]. Academy of Management Perspectives, 2021, 35(2): 265-291.
2 Ahmad I S, Bakar A A, Yaakub M R. Movie revenue prediction based on purchase intention mining using YouTube trailer reviews[J]. Information Processing & Management, 2020, 57(5): 102278.
3 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.
4 Bello-Orgaz G, et al. Marketing analysis of wineries using social collective behavior from users' temporal activity on Twitter[J]. Information Processing & Management, 2020, 57(5): 102220.
5 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.
6 Barlow M A, et al. Optimal distinctiveness, strategic categorization, and product market entry on the Google Play app platform[J]. Strategic Management Journal, 2019, 40(8): 1219-1242.

1.3 研究问题

期望解决的问题：

（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

以往研究：

产品文本 → 基础向量表示 → 相似度计算

探究：

产品文本 → **特征提取** → 相似度计算

（2）探究**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

以往研究：

价格
销量
口碑
企业规模
→ **互补者绩效预测**

探究：

与**平台所有者**的产品同质化水平
与**其他互补者**的产品同质化水平
→ **互补者绩效预测**

预测价值？促进程度？

平台所有者在平台市场中具有强主导性

中国地质大学經濟管理學院
SCHOOL OF ECONOMICS AND MANAGEMENT, CHINA UNIVERSITY OF GEOSCIENCES

2 研究设计

2.1 研究设计

研究问题：在平台强主导机制的情况下，对于互补者来说，与平台间的**产品同质化水平**是否可以成为其预测未来销量趋势的**关键指标**？

问题拆解

如何衡量产品的同质化水平?

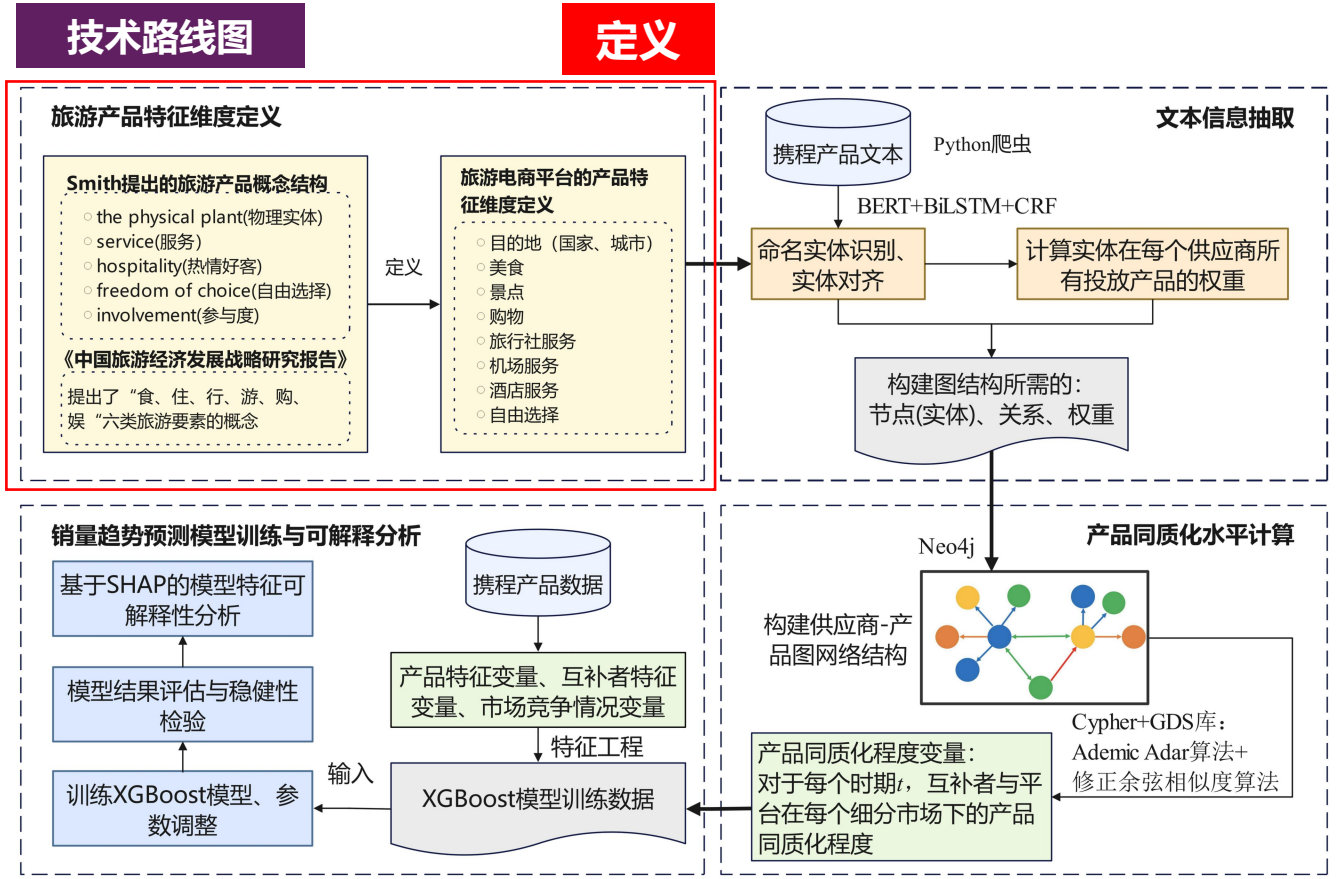
（1）产品特征的**定义、提取与表示**

（2）产品同质化水平的**计算**

如何评估产品同质化水平的预测价值?

（1）**衡量**预测价值（基于SHAP）

（2）与其他预测指标的**对比**



2.1 研究设计

研究问题：在平台强主导机制的情况下，对于互补者来说，与平台间的**产品同质化水平**是否可以成为其预测未来销量趋势的**关键指标**？

问题拆解

如何衡量产品的同质化水平?

（1）产品特征的**定义、提取与表示**

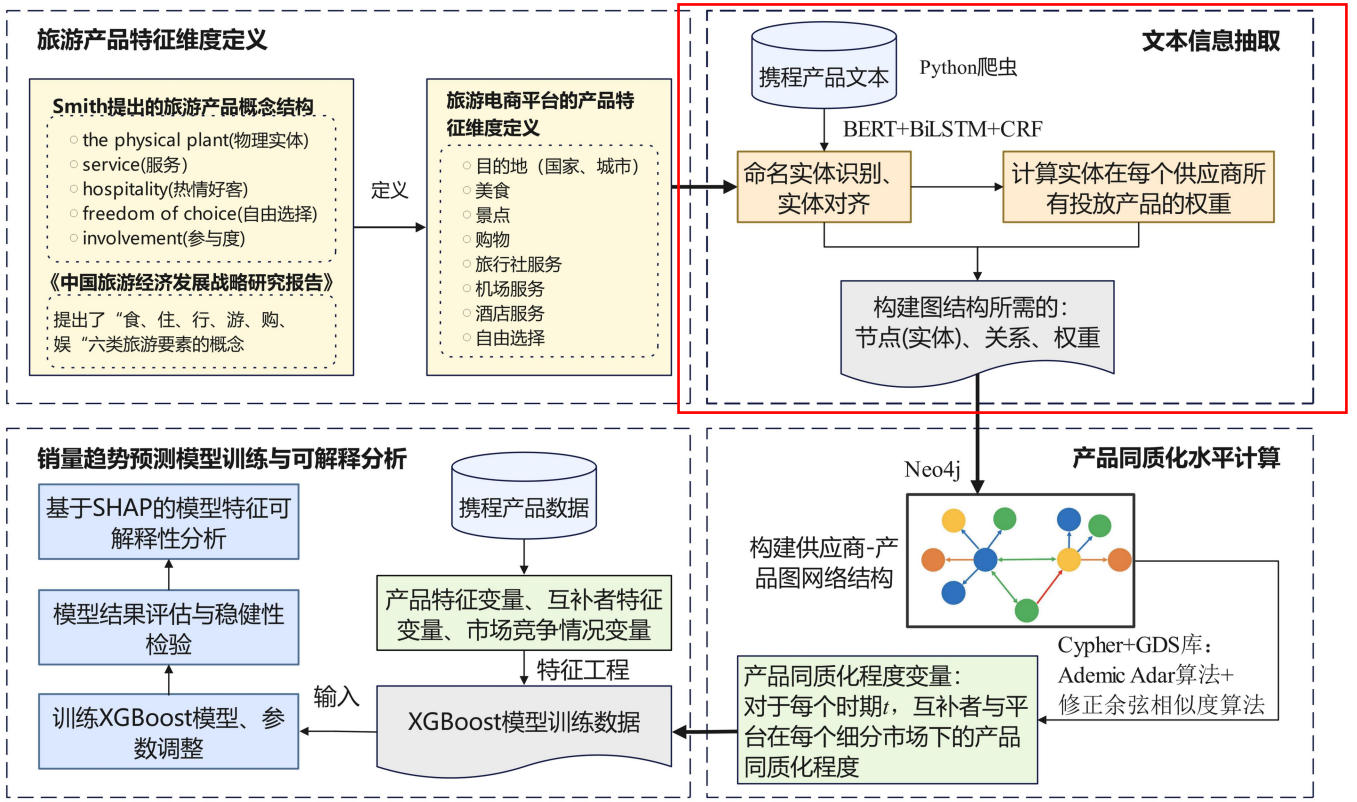
（2）产品同质化水平的**计算**

如何评估产品同质化水平的预测价值?

（1）**衡量**预测价值（基于SHAP）

（2）与其他预测指标的**对比**

技术路线图



2.1 研究设计

研究问题：在平台强主导机制的情况下，对于互补者来说，与平台间的**产品同质化水平**是否可以成为其预测未来销量趋势的**关键指标**？

问题拆解

如何衡量产品的同质化水平？

（1）产品特征的**定义、提取与表示**

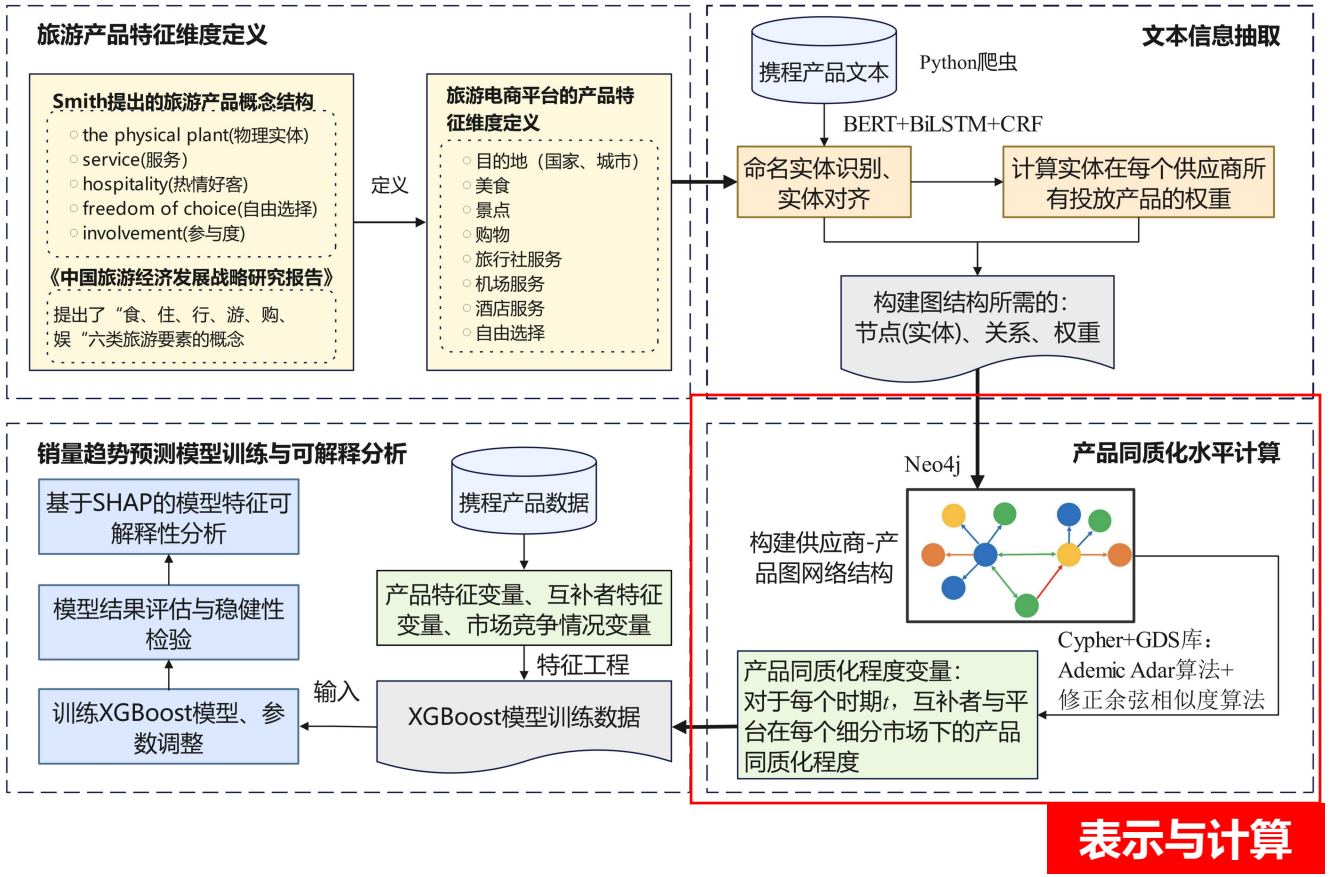
（2）产品同质化水平的**计算**

如何评估产品同质化水平的预测价值？

（1）**衡量**预测价值（基于SHAP）

（2）与其他预测指标的**对比**

技术路线图



2.1 研究设计

研究问题：在平台强主导机制的情况下，对于互补者来说，与平台间的**产品同质化水平**是否可以成为其预测未来销量趋势的**关键指标**？

问题拆解

如何衡量产品的同质化水平?

（1）产品特征的**定义、提取与表示**

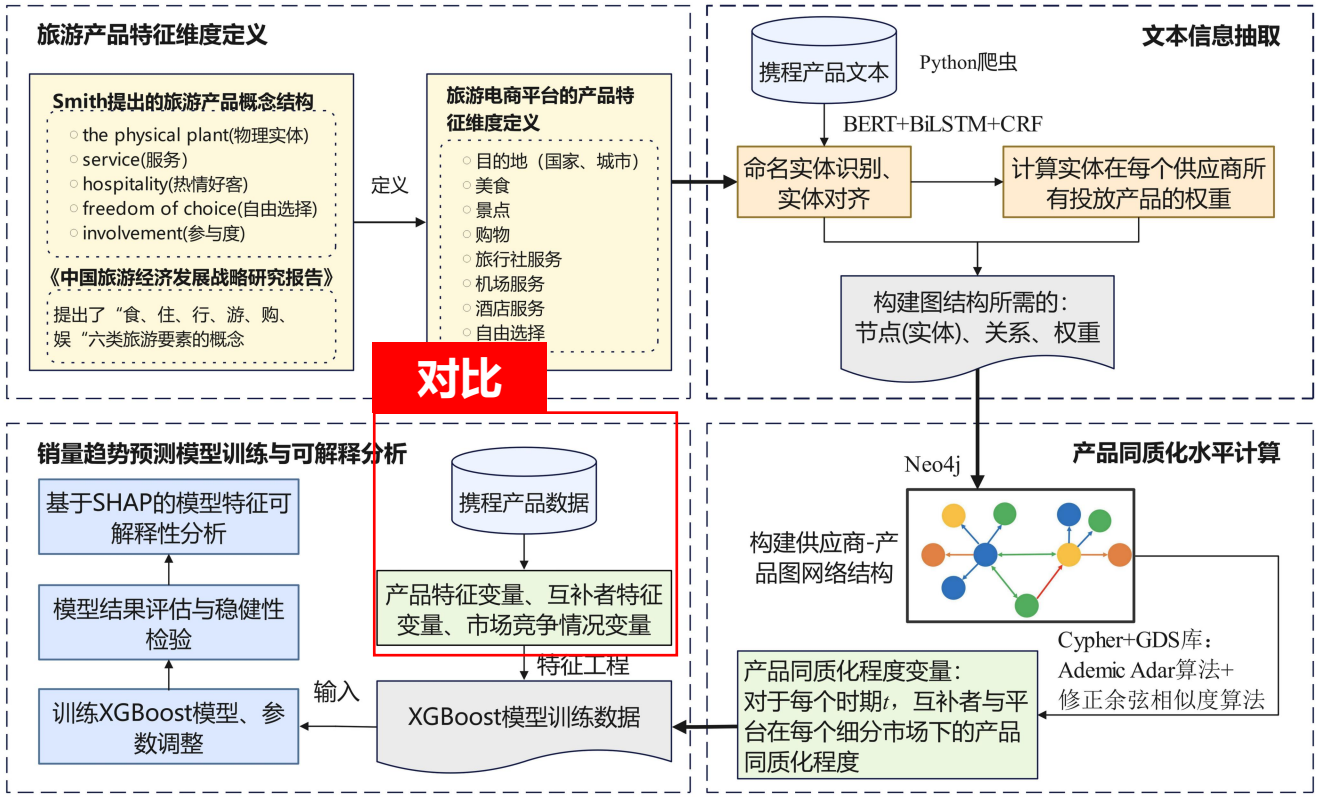
（2）产品同质化水平的**计算**

如何评估产品同质化水平的预测价值?

（1）**衡量**预测价值（基于SHAP）

（2）与其他预测指标的**对比**

技术路线图



2.1 研究设计

研究问题：在平台强主导机制的情况下，对于互补者来说，与平台间的**产品同质化水平**是否可以成为其预测未来销量趋势的**关键指标**？

问题拆解

如何衡量产品的同质化水平?

（1）产品特征的**定义、提取与表示**

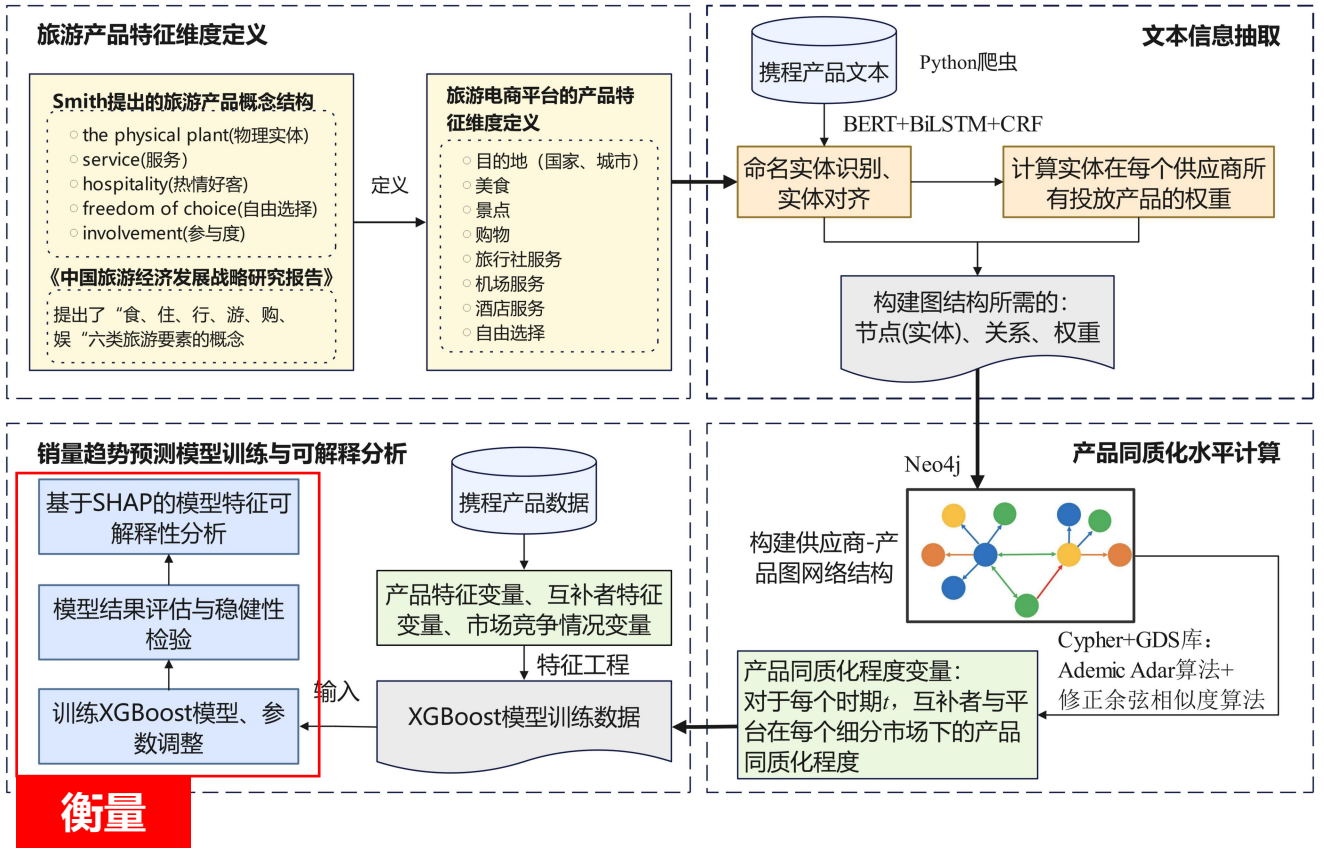
（2）产品同质化水平的**计算**

如何评估产品同质化水平的预测价值?

（1）**衡量**预测价值（基于SHAP）

（2）与其他预测指标的**对比**

技术路线图



中国地质大学經濟管理學院
SCHOOL OF ECONOMICS AND MANAGEMENT, CHINA UNIVERSITY OF GEOSCIENCES

3 数据与变量测量

- 数据来源
- 产品同质化水平的计算
- 变量定义

3.1 数据

数据来源 Ctrip 携程

国内**头部**旅游电商平台，提供全球旅行服务

截至2021年，携程的市场份额约为**70%**。

数据爬取范围

共**42252**个产品

时间：按月爬取，**2017年3月至2018年12月**

出发城市与目的地国家：北京 → 美国

- 产品**多元**、市场**规模大**
- 出境游的产品**内容丰富**、对供应商**资质要求更高**

爬取内容：产品id、标题、服务保障、供应商、产品卖点、价格、销售量、评论量、评分等数据。

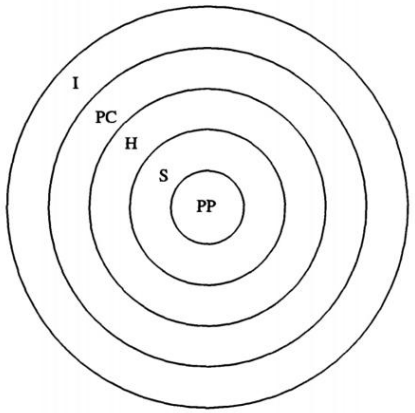
携程平台上的产品页面[1]

1 图片来自 https://hotels.ctrip.com

3.2 产品同质化水平的计算

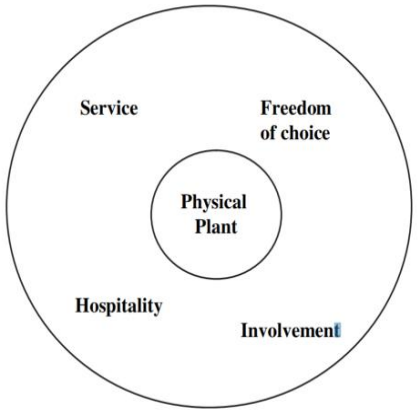
产品特征的定义：基于Smith提出的旅游产品五大组成元素，本研究定义了**九大维度**表示旅游电商产品

Smith和Xu先后提出的旅游产品概念化定义：



PP = Physical Plant　　FC = Freedom of Choice
S = Service　　I = Involvement
H = Hospitality

（Smith，1994）



（Xu，2010）

本研究提出的旅游电商产品的九大维度特征定义

旅游产品五大组成元素	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	/

1 Smith S L J. The tourism product[J]. Annals of tourism research, 1994, 21(3): 582-595.
2 Xu J B. Perceptions of tourism products[J]. Tourism management, 2010, 31(5): 607-610.

3.2 产品同质化水平的计算

产品特征的抽取：基于BERT-BiLSTM-CRF命名实体识别框架的实体抽取

特征抽取步骤：（技术细节在附录A展示）

文本标注

- 使用BIO方法标注
- 共标注了3000条数据，随机抽样检查的正确率在90%以上

实体抽取

- 使用BERT+BiLSTM+CRF框架抽取

实体合并

- 实体删除
- 实体对齐

模型抽取结果示例：

产品标题文本
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

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

3.2 产品同质化水平的计算

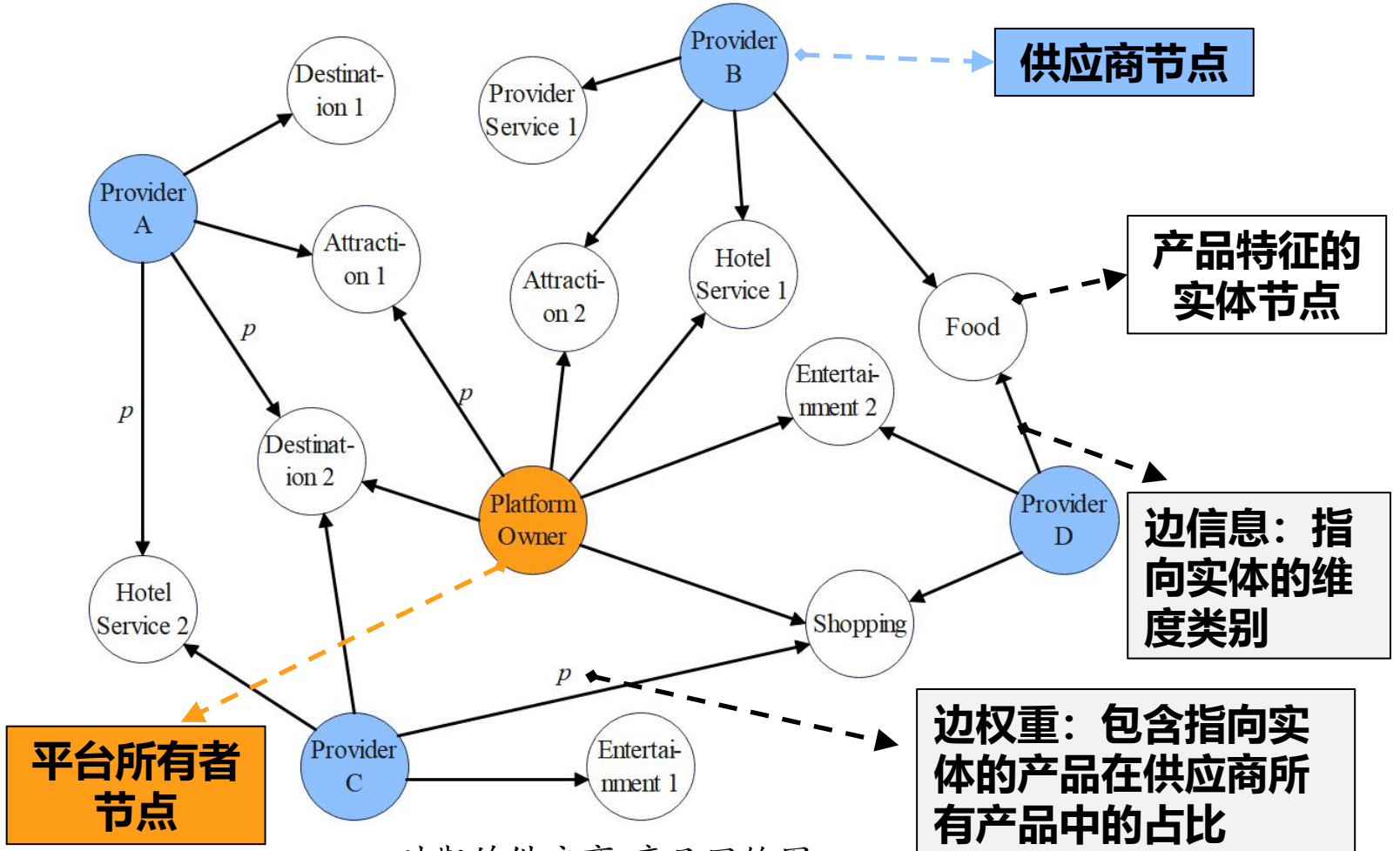
供应商产品特征的表示：构建供应商-产品图网络

单个产品特征的表示

▼

聚合单个产品：供应商的整体产品投放特征表示

细节：
在具体构建图网络的过程中，由于之后要服务于供应商之间的产品同质化水平计算，本研究进一步将美国旅游市场拆分为四个细分市场，并针对每个细分市场的每个时期都构建了一个图网络。



*t*时期的供应商-产品网络图

3.2 产品同质化水平的计算

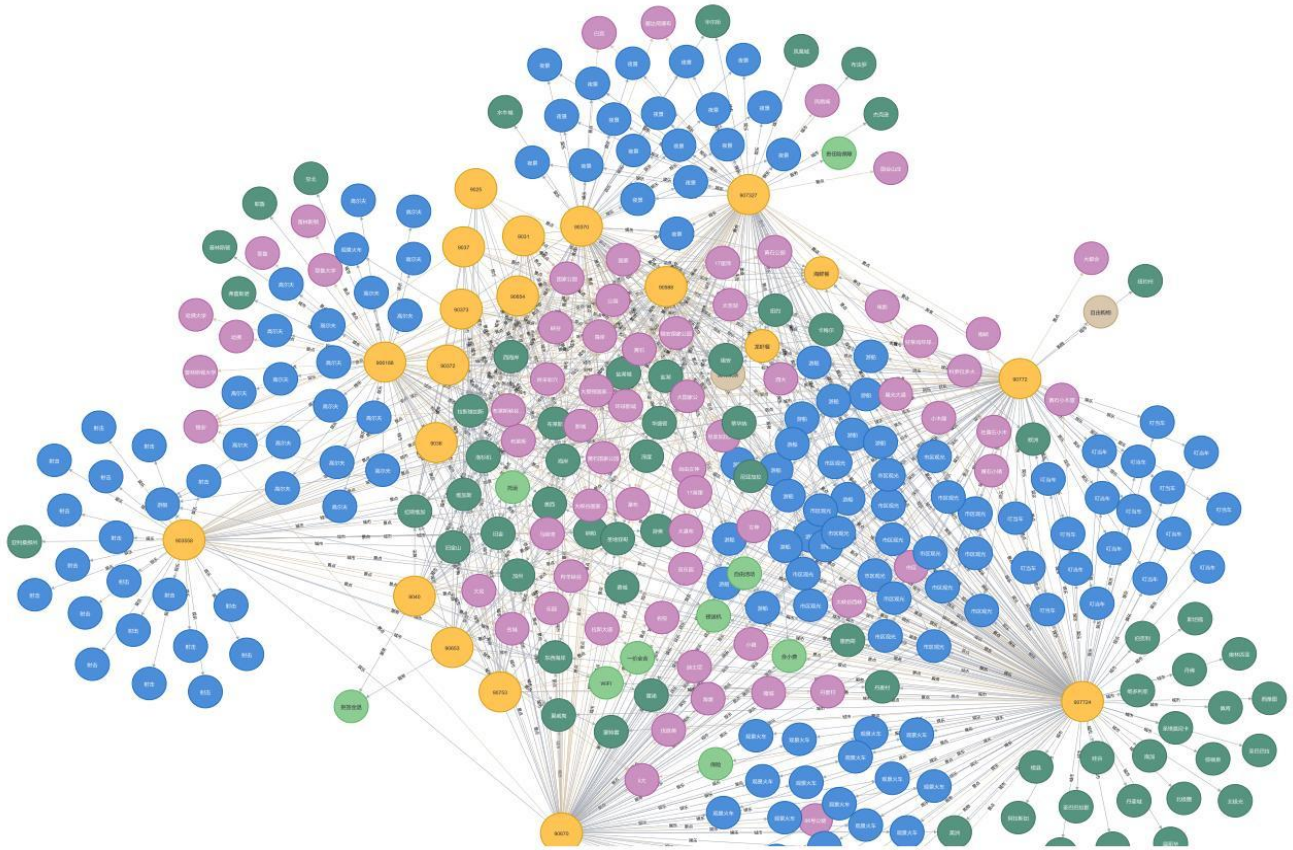
供应商产品特征的表示：构建供应商-产品图网络

单个产品特征的表示

▼

聚合单个产品：供应商的整体产品投放特征表示

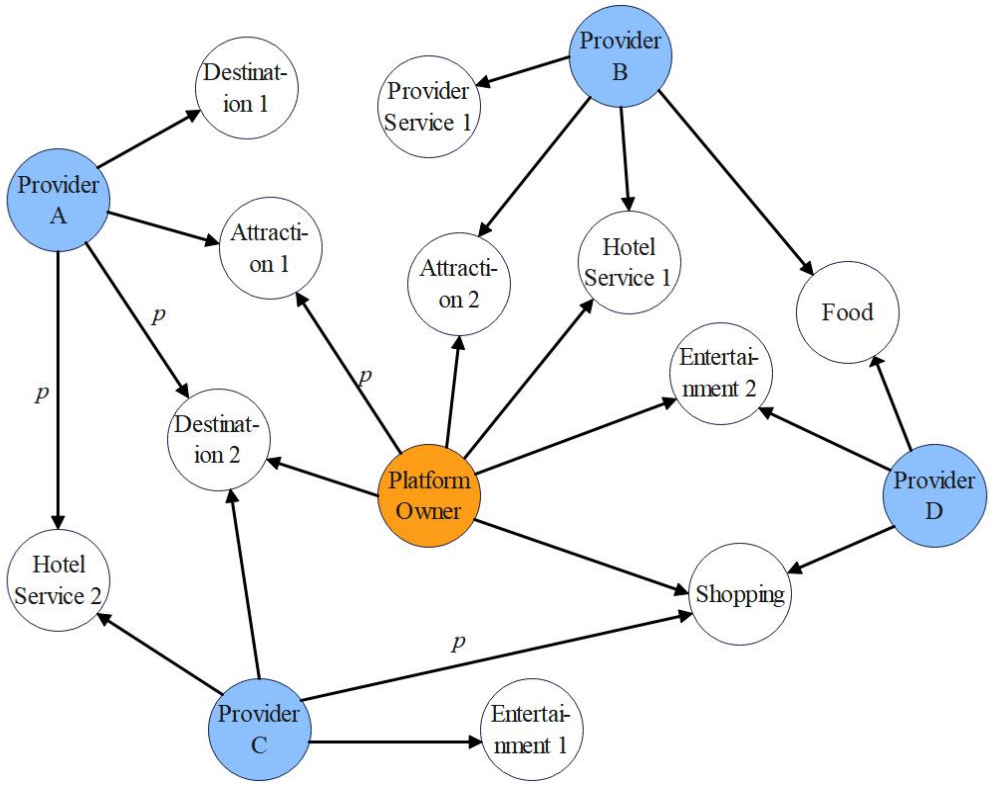
细节：
在具体构建图网络的过程中，由于之后要服务于供应商之间的产品同质化水平计算，本研究进一步将美国旅游市场拆分为四个细分市场，并针对每个细分市场的每个时期都构建了一个图网络。



以旧金山细分市场为例的供应商-产品网络Neo4j可视化展示（2017年12月）（附录C）

3.2 产品同质化水平的计算

互补者与平台间产品同质化水平的计算：图节点相似度计算



*t*时期的供应商-产品网络图

Ademic-Adar算法：

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

计算两个节点间**公共邻居节点**的重合程度

+

修正的余弦相似度算法：

$$sim(i,j)=\frac{\sum_{u\in U}(R_{u,i}-\overline{R_u})(R_{u,j}-\overline{R_u})}{\sqrt{\sum_{u\in U}(R_{u,i}-\overline{R_u})^2}\sqrt{\sum_{u\in U}(R_{j,i}-\overline{R_u})^2}}$$

计算两个节点间连接的每个公共邻居节点的**权重系数**相似程度

=

供应商间产品同质化水平的计算方法：

$$S(x,y)=A(x,y)\times\frac{sim(W_x,W_y)+1}{2}$$

值域映射到(0,1)之间

3.3 变量定义

本研究重点关注产品同质化水平这一指标对互补者销量趋势预测模型的贡献价值。因此，额外加入了以下三类变量作为评估产品同质化水平变量预测价值的**对比**。

变量分类：

一、互补者的特征

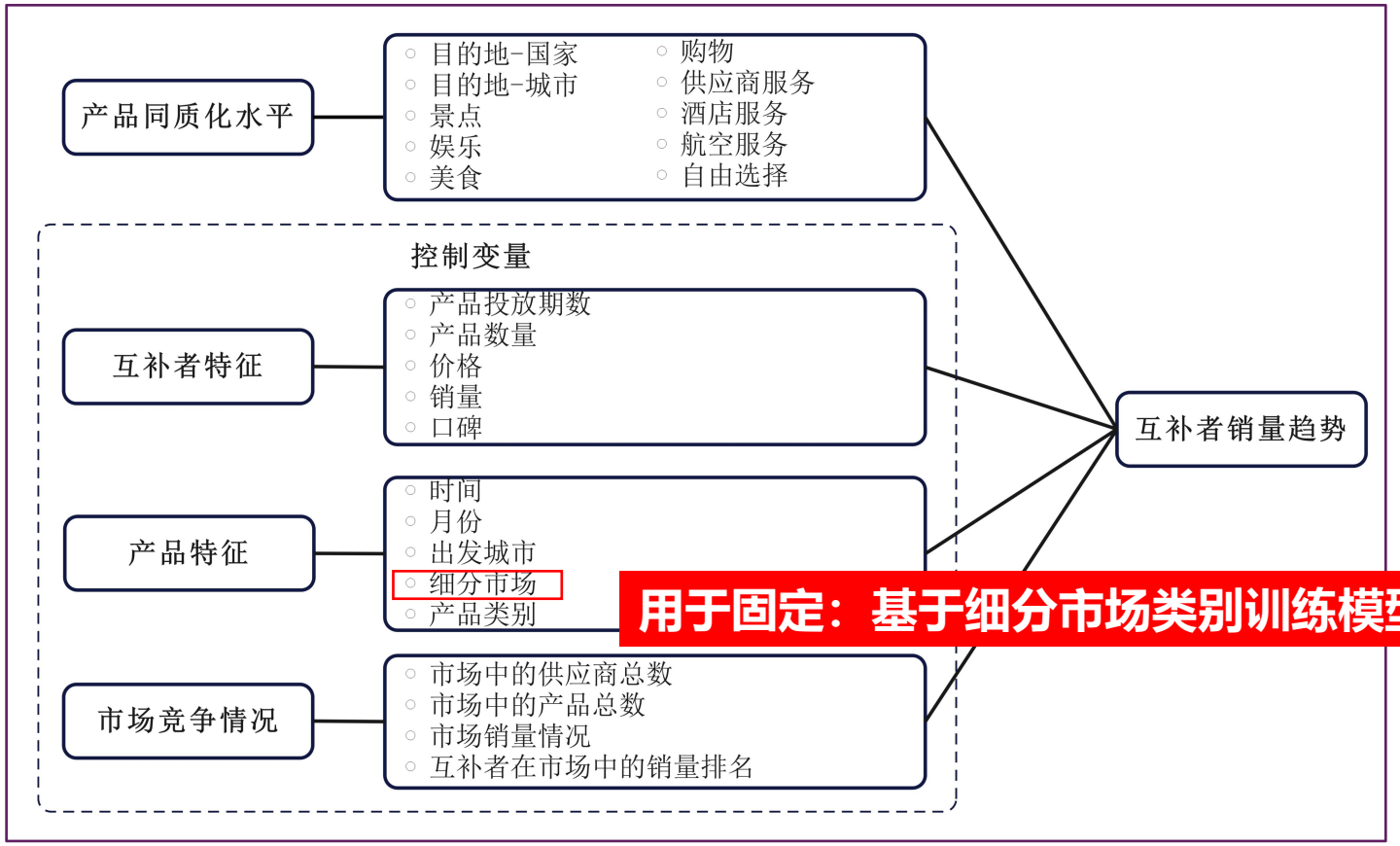
与互补者相关的随时间变化的变量，反映互补者内在特征对销量趋势预测的影响。

二、市场竞争情况

与互补者无关的随时间变化的变量，反映某个时期内市场环境的竞争程度。

三、产品特征

与互补者和时间都无关的变量，在本实验中，该类变量后期主要用于划分具体的细分市场以计算产品同质化水平。



中国地质大学經濟管理學院
SCHOOL OF ECONOMICS AND MANAGEMENT, CHINA UNIVERSITY OF GEOSCIENCES

4 预测模型构建

4.1 数据集预处理、模型选择与训练

数据集预处理

时间变量的处理： 日期（2017年3月至2018年12月） ▶ 数值变量（1-22）

目标值Y的设定： （0-1）互补者在下一期销量是否上升？
- 1 上升
- 0 下降或持平

模型选择与训练

本研究的销量趋势预测模型基于**XGBoost模型**建立，原因如下：
（1）在高维数据集中表现良好；
（2）贪婪算法，选择信息成分最大的特征进行预测；
（3）灵活处理预测因子间的潜在相关性。

XGBoost与其他模型的预测结果对比（AUC）

	XGBoost	Random forest	SVM
Baseline	0.761	0.735	0.713
Baseline+similarity	0.779	0.749	0.730

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

4.2 模型结果评估

与基线模型对比：

Baseline模型：

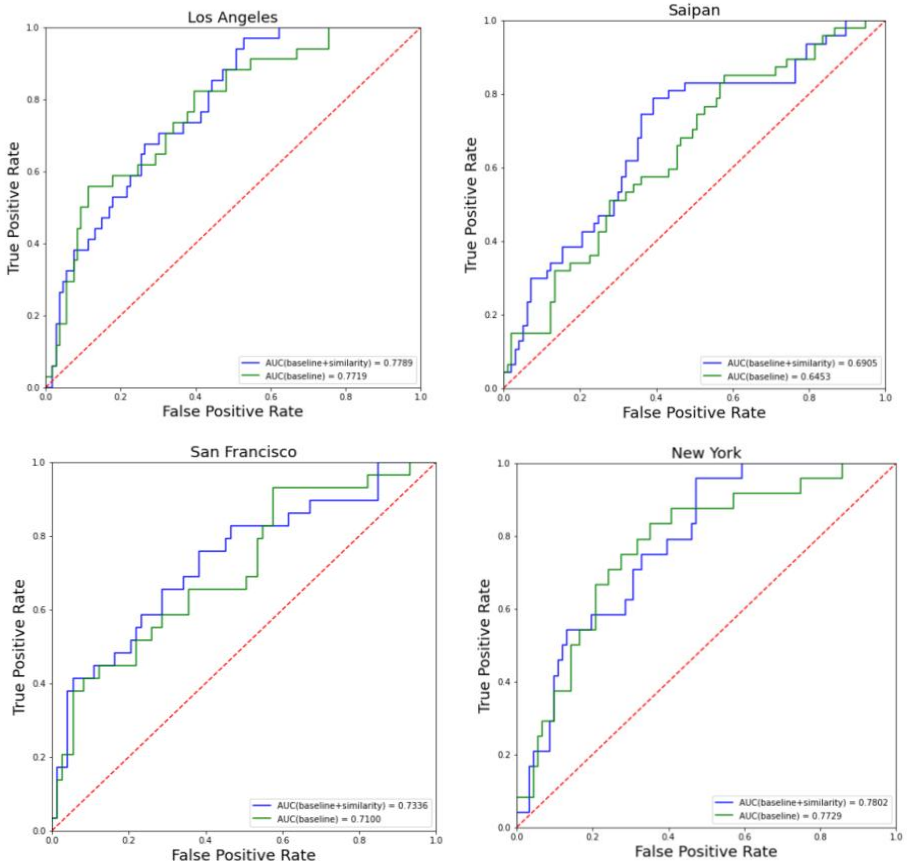
互补者特征、产品特征、市场竞争情况

Baseline+Similarity模型：

互补者特征、产品特征、市场竞争情况 + 产品同质化水平

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



模型结果对比（ROC曲线）

4.3 模型稳健性检验

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：

1. **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；
2. **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平台间产品同质化水平的算法，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

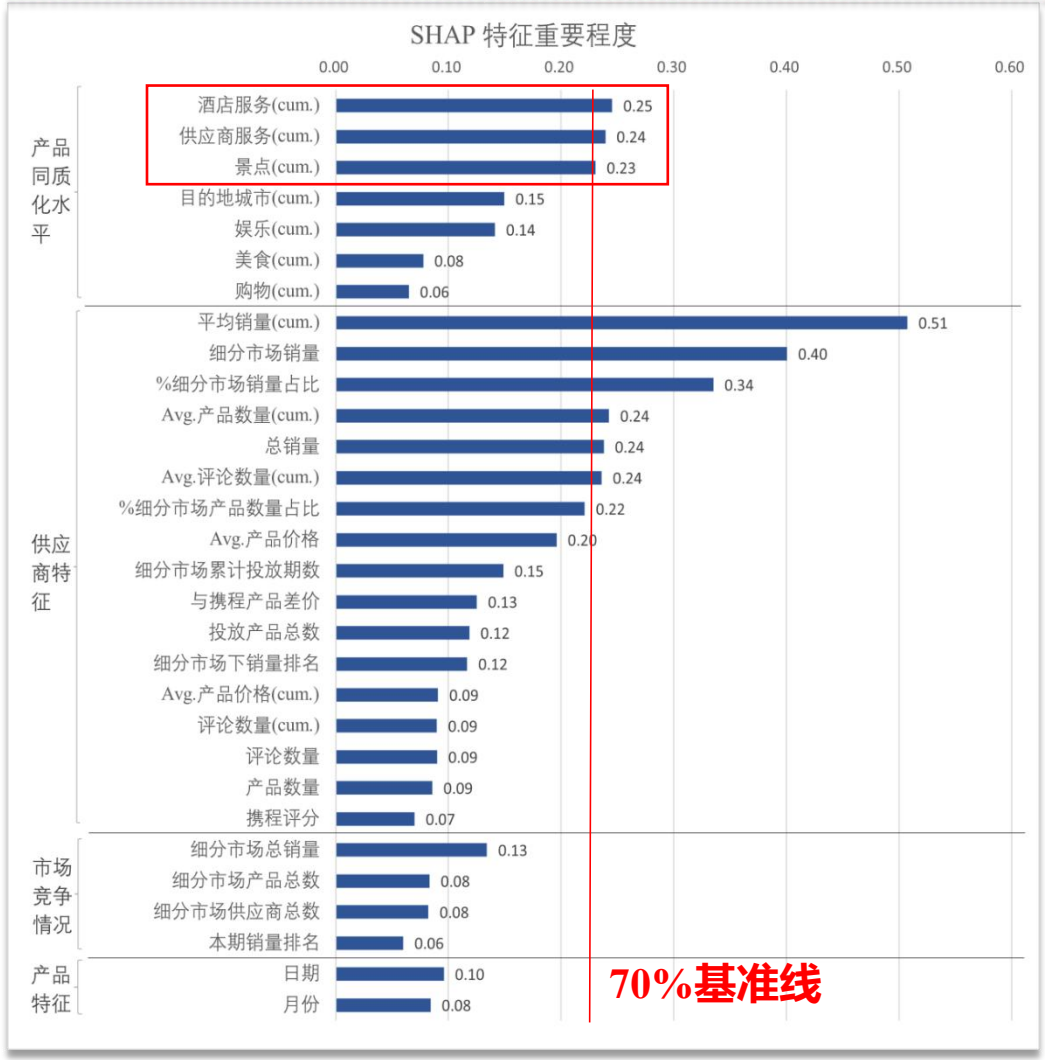
稳健性检验方式	模型（整体）	评价指标					
		AUC	Accuracy	Precision	Recall	F1	KS
替换累计变量	Baseline	0.6649	0.7662	0.6000	0.3000	0.4000	0.2298
	Baseline+Similarity	0.6721	0.7549	0.6000	0.4138	0.4898	0.3042
替换产品相似度	Baseline	0.6535	0.7033	0.6087	0.3373	0.4341	0.2269
	Baseline+Similarity	0.6798	0.7410	0.5393	0.3529	0.4267	0.2397

中国地质大学经济管理学院
SCHOOL OF ECONOMICS AND MANAGEMENT, CHINA UNIVERSITY OF GEOSCIENCES

5 基于SHAP的模型可解释性分析与研究结论

研究结论1：特征重要性分析

产品同质化水平在互补者销量趋势预测中起到了重要作用，其特征重要程度超过了70%的变量



重要性程度最高的**前30个变量**
（四个模型均值）

1. 最具贡献价值指标：**平均销量、细分市场整体销量和细分市场销量占比**
2. 贡献价值：**累计平均指标** > 单期指标
3. 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**这三个维度的同质化程度指标对模型的贡献也较大，超越了**70%**的特征。

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。

研究结论2：特征相关性分析

细分市场的产品内容中占比越高的特征，在销量趋势预测中作用越大

对目的地城市维度进行**聚类**：

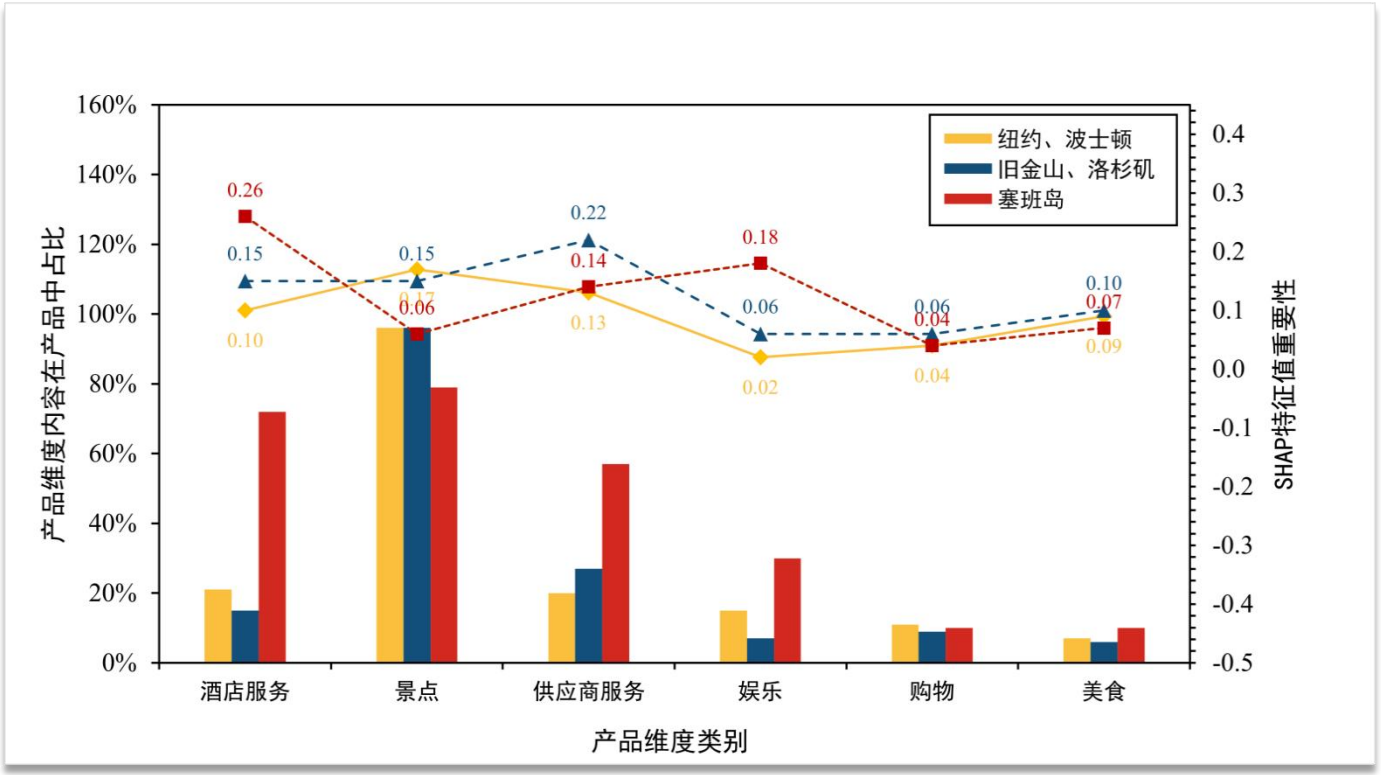
纽约
旧金山
洛杉矶
塞班岛

合并频繁共现的城市

K-Means聚类

纽约、波士顿
旧金山、洛杉矶
塞班岛

细分市场类别 ⟺ 相关性 ⟺ **产品各维度的预测价值**



当含有某维度特征的产品在该市场所有产品中**占比越高**时，该维度特征的预测**价值越大**。

产品的每个特征在预测中的价值大小并不是绝对的，与该细分市场的产品特点有关。

研究结论3：特征相关性分析

与平台间的产品同质化水平指标对产品规模较小的互补者的销量趋势预测价值更大

对目的地城市维度进行**聚类**：

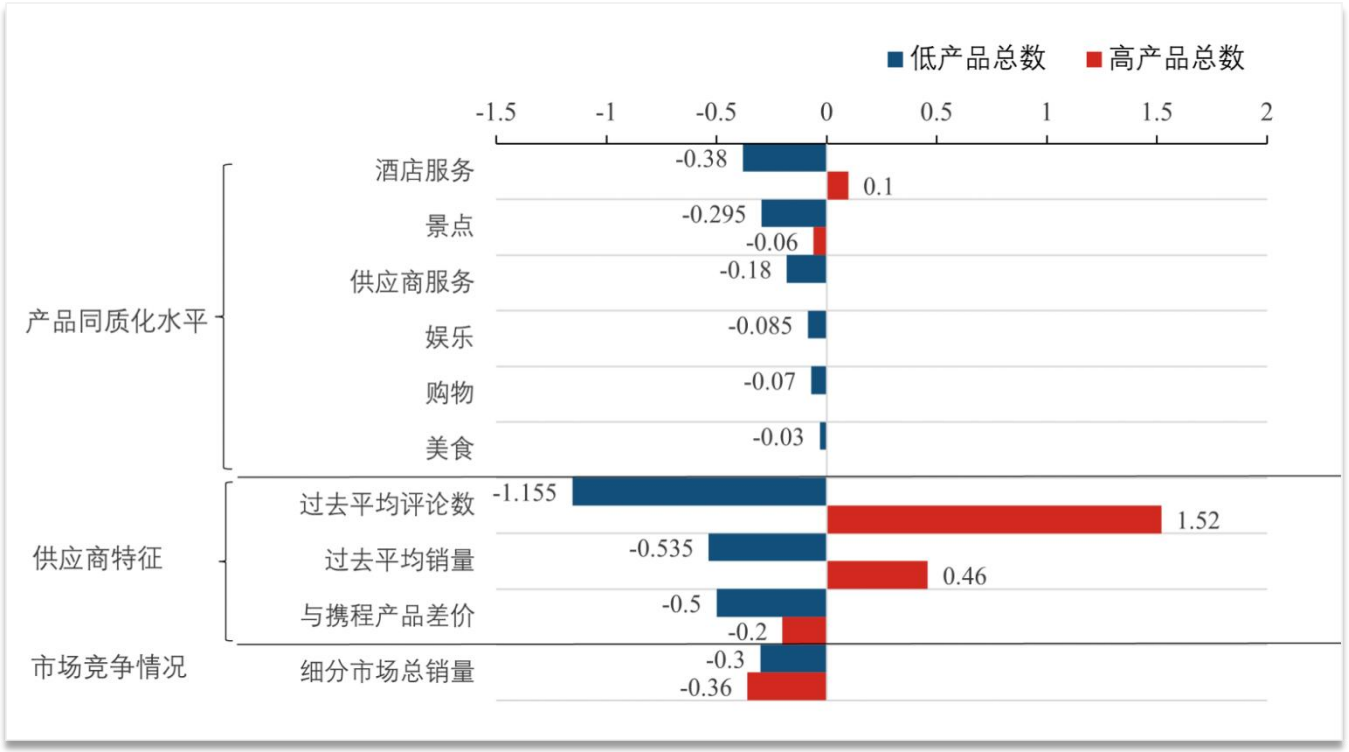
纽约
旧金山
洛杉矶
塞班岛

合并频繁共现的城市

K-Means聚类

纽约、波士顿
旧金山、洛杉矶
塞班岛

互补者产品规模 ⟷ 相关性 ⟷ **产品各维度的预测价值**



对产品规模**较小**的互补者，与平台间产品同质化水平的指标具有**更大**的预测价值，且这些指标与预测结果呈现**负相关**关系

反映特征值与预测结果之间的**正负相关性**

6 不足与展望

6.2 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，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。未来可以使用因果森林等因果识别方法，进一步探究**因果效应**以及特征的**具体影响程度**。

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。

答辩结束

请各位老师批评指正！

中国地质大学（武汉） 学士学位毕业论文答辩
电商平台中互补者的产品同质特征与销量趋势预测研究
指导教师：朱镇 教授　　答辩人：徐嘉艺