目录

《数据挖掘与商务分析》期末报告

星图·言迹·群荐——基于文本、时空与社交的商户智能推荐系统

小组成员信息:

学号	姓名	工作分工	成绩
2022111155	买海成	整体流程设计，流程图生成；负责时空聚类、文本情感分析；完成论文摘要，第一、二、五、六、七部分的撰写	33.3%
2022111051	李乔鑫	参与整体流程设计，负责社交网络构建以及基于社交网络的推荐策略分析，论文负责第二部分和第五部分的社交网络分析	33.3%
2022110968	翟誉钧	参与整体流程设计，进行数据预处理与 xgboost 预测，完成论文第三第四部分	33.3%

完稿日期: **2025 年 6 月 29 日**

星图·言迹·群荐——基于文本、时空与社交的商户智能推荐系统

摘要

本研究围绕 Yelp 商户数据，构建了一个结合空间、文本与社交信息的新商户推荐系统。首先，基于商户结构化属性、地理坐标、类别与名称文本等信息，构建特征矩阵并利用 XGBoost 回归模型实现对商户星级（Stars）的预测，验证集平均绝对误差（MAE）为 0.57。其次，设计了试运营期间的主观评价机制，借助情感分析工具对评论文本进行预处理和极性评分，提出“好评率”指标，用于辅助识别高质量新商户。接着，从用户-评论数据中构建社交网络图，采用滚雪球抽样方法提取核心子图，通过中心性分析识别种子用户，并使用 Louvain 算法进行社群识别，形成基于社交结构的精准推荐策略，实现“优质商户优先推荐给其评论者的朋友”。综合上述模型与流程，最终形成一套融合地理分布、文本情感与用户社交图谱的复合型推荐系统，面向真实应用场景提供智能化、可解释的新商户推广解决方案。研究还探讨了模型性能、特征构造、社交连接等方面的局限性，并展望未来通过引入深度学习模型和协同推荐算法进一步提升系统的预测精度与推广效果。

Abstract

This study presents an integrated recommendation system for new businesses on Yelp, incorporating spatial, textual, and social information. First, a predictive model is developed using XGBoost to estimate the star ratings of businesses based on structured attributes, geolocation, and TF-IDF features derived from category and name text, with a validation Mean Absolute Error (MAE) of 0.57. Next, to simulate early-stage subjective feedback, sentiment analysis is applied to review texts using polarity scoring. A “positive review rate” is computed by identifying reviews with polarity > 0.2, serving as a complementary indicator of business quality during trial operations. To enhance recommendations via social context, a user social network is constructed based on the “friends” field. A snowball sampling strategy is applied to extract a representative subgraph, followed by centrality-based seed user identification and community detection using the Louvain algorithm. High-quality businesses are then recommended to friends of early reviewers, leveraging network influence for effective promotion. The final system combines spatiotemporal features, textual sentiment, and user graph structure to deliver a multi-perspective, explainable recommendation strategy. Limitations such as feature sparsity, polarity threshold sensitivity, and cold-start gaps in user-business matching are discussed, with future work proposed to include collaborative filtering and deep learning models for temporal sentiment forecasting.

一、项目背景及意义

1.1 项目背景

随着本地生活服务平台的快速发展，用户生成内容（User-Generated Content, UGC） 和社交网络在消费者决策中的影响力日益增强。UGC，特别是用户对商户的评论文本和评分数据，为平台提供了大量关于消费者偏好的主观反馈。然而，这些信息往往受到用户情绪、个体认知等因素的影响，存在一定程度的主观误差。过去的研究中，许多模型主要依赖评论文本进行星级预测，侧重于挖掘情感倾向和语言特征[1]，却忽视了商户自身的**客观属性（如营业时间、种类、地理位置、服务特色等）**在用户评分决策中的关键作用。

在此背景下，如何将用户主观评价与商户客观属性相结合，开展多源数据融合分析，成为实现更全面、准确商户画像与用户行为建模的关键。通过结合机器学习、自然语言处理和图分析等方法，可对消费者偏好与商户特征之间的关系进行系统性建模，从而提升平台在商业洞察、个性化推荐与城市消费引导等方面的智能决策能力。

本研究以 Yelp 平台提供的多维度开源数据为基础，涵盖商户基本属性、用户评分与评论文本、用户社交关系等信息，探索从主客观数据融合、时空聚类建模到社交网络分析等多视角展开的数据挖掘方法，具有重要的研究意义与应用价值。

1.2 项目意义

本项目基于 Yelp 多源异构数据，围绕商户星级评分的影响因素，综合运用自然语言处理、空间聚类、社交网络建模与机器学习方法，旨在从多个维度深入挖掘用户行为与商户特征之间的关系，从而提升模型的**解释力、预测力与推荐能力。**

文本挖掘层面，本研究利用自然语言处理技术对用户评论文本进行情感分析，并引入“好评率”指标，量化用户评价的整体情绪倾向。同时结合 LDA 等主题建模方法，对商户类别进行建模。有助于揭示用户评分背后的体验维度与关注重点，从而增强星级预测模型的可解释性。

空间分析层面，基于商户的地理位置信息，通过 K-Means 聚类方法构建城市消费聚集区，引入“便利性”指标，衡量商户与聚类中心的空间距离，以探讨地理区位对商户评分和用户偏好的影响。此外，结合城市空间结构，可进一步提出商户选址与布局的优化建议。

社交网络分析层面，通过构建用户间的社交图谱，模拟信息或偏好在用户之间的传播路径，提出“好店优先推荐给朋友”的个性化推荐策略，提升模型的可扩展性与用户粘性。

预测建模层面，整合地理位置、便利性、商户类型等多源特征，构建基于 XGBoost 的星级评分预测模型，显著提升模型性能。该方法可为本地生活服务平台和餐饮行业提供数据驱动的精准营销与运营决策支持。

综上，本项目不仅实现了对 Yelp 商户评分的多角度建模与解释，也为本地生活服务平台在用户推荐、服务优化和智能决策等方面提供了实证支持与方法参考。

二、问题描述

图1：主题视角下项目流程图

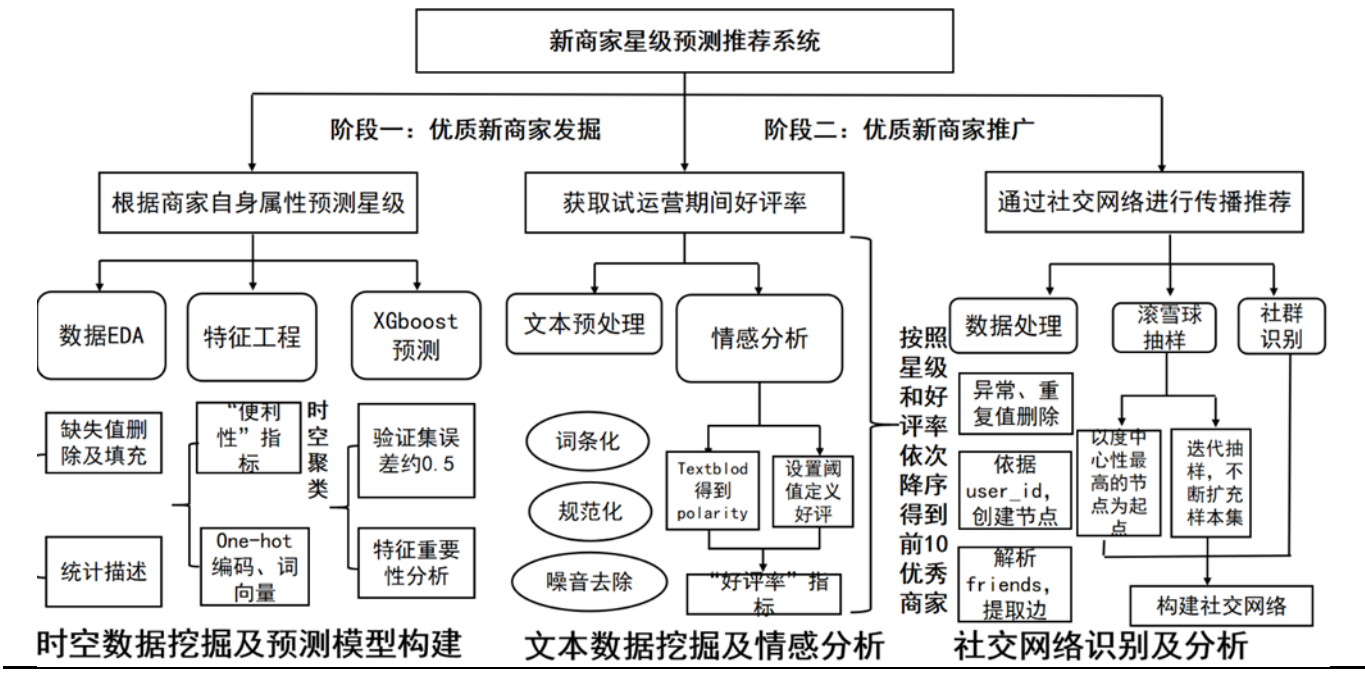
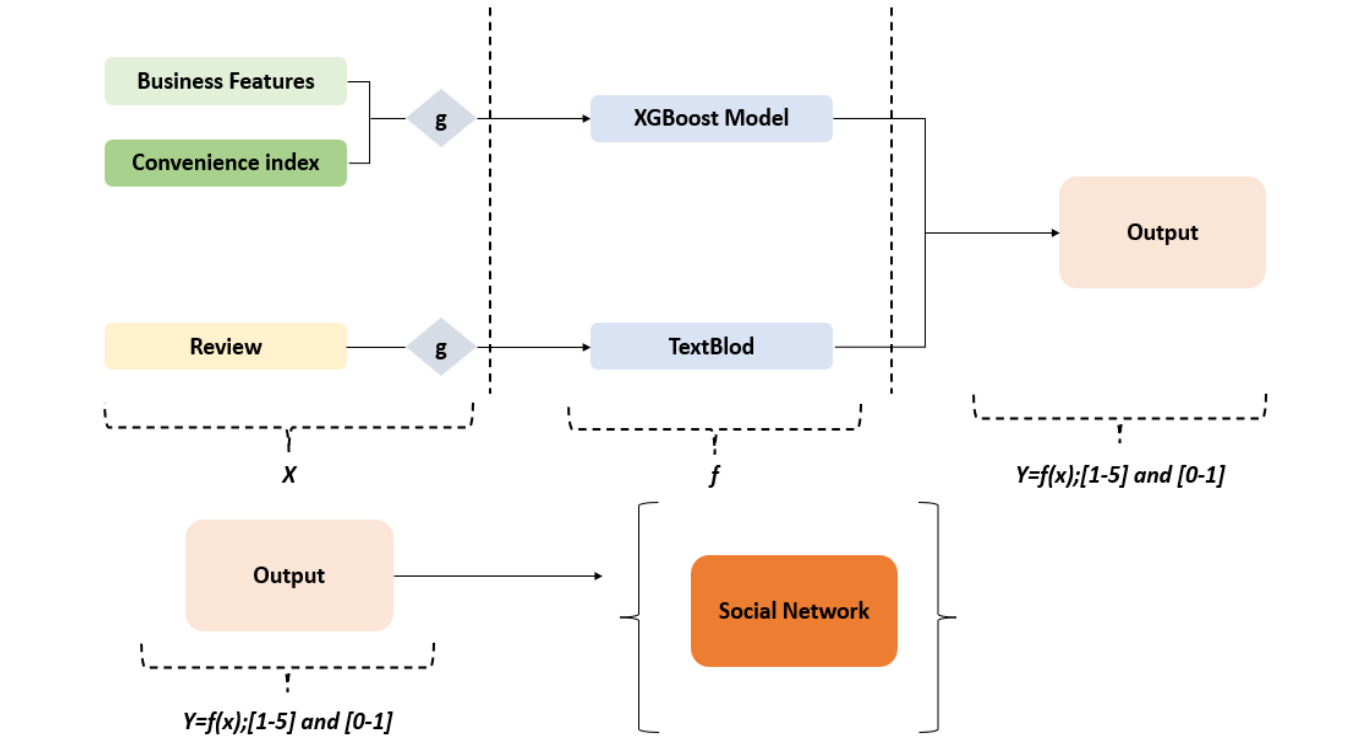


图2：机器学习数据挖掘视角下项目流程图



本项目基于 Yelp 多源数据，围绕“商户评分预测—试运营情感挖掘—社交网络推荐”三大核心问题展开建模，结合 XGBoost 回归、文本情感分析、空间聚类与社交网络传播等方法，形成了系统化的数据挖掘流程。具体包括以下五项分析任务：

（1）基于商户属性预测新商户的评分水平

项目首先关注新商户在无用户评价的“冷启动”阶段的评分预测问题。通过提取 Yelp_business.json 中的商户类别、营业状态、地理位置等结构化特征，并结合商户与城市聚类中心之间的距离构建“便利性”指标，从而形成特征矩阵用于建模。

所采用模型为 XGBoost（eXtreme Gradient Boosting），[2][3]其核心思想是在 Boosting 框架下迭代训练多棵 CART（分类回归树），通过加权组合预测结果并最小化目标损失函数。

（2）基于评论文本的试运营情感挖掘

项目进一步尝试衡量商户试运营阶段的主观表现，使用 reviews_of_restaurants.txt 中的评论数据对每个商户进行聚合（group by business_id），并筛选验证集中每家商户最多前 50 条评论（按时间排序）作为其“试运营评论集”。

文本分析中使用了 TextBlob（textplod）库中的 情感分析器（Naive Bayes + polarity 规则），[7]对每条评论文本提取 polarity 分数（范围 -1~1），代表情感倾向。根据经验阈值，设定 polarity > 0.2 判定为“好评”，进而计算该商户试运营期间的“好评率”。

该方法实质上实现了从非结构化文本中提取结构化情绪指标，为后续评价解释与商户筛选提供支持。

（3）融合评分与好评率进行优质商户识别排序

项目综合考虑模型预测评分（XGBoost 输出）与用户主观好评率两个维度，对所有新商户进行联合排序，按照“预测评分 + 好评率”依次降序排列，评选出前 10 位表现最优的试运营商户。

这一方法本质上是一种多指标融合排序机制（score-ranking aggregation），兼顾了主客观信息的可信度与互补性，有利于从多维度识别真正潜力商户。

（4）基于用户社交关系网络建模

项目进一步将推荐系统拓展至用户的社交网络层面，利用 user.txt 中的 friends 字段构建用户之间的无向图（节点为 user_id，边为好友关系对），共解析出约 36 万条用户关系。为控制网络规模与计算复杂度，我们采用滚雪球抽样策略，从度中心性最高的用户作为初始节点出发，逐步扩展其邻居节点，构建最大深度为 5、最大节点数为 10000 的子图。

该子图可更有效捕捉社交传播结构，并显著减少冗余节点对推荐系统的干扰。通过对这一社交子图的结构分析，我们为每个用户建立其在社交网络中的上下游传播路径，为后续“优质商户信息如何更有效传递”提供了基础框架。整个过程体现了社交推荐系统（Social Recommender System）的核心思想——利用用户之

间的关系网络提升推荐精准性与传播广度。

（5）评估社交用户的传播能力并进行社群识别

为了进一步提升社交传播效率，项目计算了子图中每个用户的度中心性与介数中心性，用于评估其在推荐中的影响力。其度中心性（Degree Centrality）衡量了用户节点的直接连接数，用以识别在网络中最具连接能力的活跃用户；而介数中心性（Betweenness Centrality）衡量了用户在其他用户最短路径中出现的频率，体现其在信息传播路径中的桥梁作用。

结合这些指标，我们识别出在社交传播中最具代表性和影响力的“关键节点用户”，即“种子用户”。在推荐策略中优先面向这些用户进行信息投放，有助于在最短路径内将优质商户推荐内容扩散至更大范围的用户群体，从而增强推荐系统的整体传播效率与影响力。

此外，我们也运用了 Louvain 算法进行社群识别，以进一步细化用户群体并优化传播策略。Louvain 算法通过优化模块度（Modularity）来识别网络中的社区结构，其核心思想是将网络划分为若干个内部连接紧密、外部连接稀疏的社区。借助社群识别，我们能够划分出具有相似兴趣或行为模式的用户群体，从而针对不同社群制定差异化的推荐策略。

上述五项任务共同构成了一个以预测分析、情感建模与社交推荐为核心的端到端多模态建模系统。项目不仅解决了冷启动情境下的商户评分预测问题，也通过社交传播机制实现了信息的个性化扩散，体现了数据挖掘方法在本地生活服务平台中的实际应用价值。

三、数据集介绍

3.1 数据来源

本研究数据取自 Yelp 于 2024 年发布的最新版 Yelp Open Dataset。该公开数据集涵盖 2004–2023 年间北美多个主要城市的本地商户信息、用户画像与评论记录，具有时间跨度长、行业覆盖全、标签完备等特点。为了保证区域经济环境和消费文化的一致性，本文进一步从完整数据集中仅抽取位于宾夕法尼亚州费城（Philadelphia, PA）及其都会区的商户及对应用户-评论记录。该空间过滤操作使得研究对象在地理和社会层面更具可比性，也避免了跨城市异质性对模型估计的干扰。经过筛选，得到的费城子集仍然保留了足够的样本规模和多样性，满足后续情感分析与评分机制研究的需求。

3.2 数据表结构与表间关系

3.2.1 yelp_business.csv 字段说明

列名	数据类型*	说明
business_id	string	商户唯一标识（主键）
name	string	商户名称
address	string	详细街道地址
city	string	所在城市
state	string	州 / 省份二字母缩写
postal_code	string/int	邮政编码
latitude, longitude	float	WGS-84 坐标，用于地图可视化或距离计算
stars	float	Yelp 平均星级（1–5，支持半星）
review_count	int	评论总数
is_open	int (0/1)	营业状态；1 = 正常营业，0 = 已关闭或停业
attributes	string (JSON-like)	以花括号包裹的键值对，内含“是否有外卖”“Wi-Fi”等店铺属性
categories	string	多个逗号分隔的行业标签（如 Restaurants, Brewpubs）
hours	string (JSON-like)	各星期营业时间，键为星期，值为“HH:MM-HH:MM”

3.2.2 reviews_of_restaurants.txt 字段说明

列名	数据类型	说明
review_id	string	评论唯一标识（主键）
user_id	string	评论者 ID，外键指向 users.txt
business_id	string	被评论的商户 ID，外键指向 yelp_business.csv
stars	float	此条评论给出的星级
date	datetime	评论时间（UTC-8，Yelp 默认时区）
useful	int	其他用户标记的 “useful” 票数
text	string	评论正文，原始自然语言文本

3.2.3 users.txt 字段说明

列名	数据类型	说明
user_id	string	用户唯一标识（主键）
name	string	Yelp 昵称
review_count	int	用户累计发布的评论数量
yelping_since	datetime	注册时间
friends	string / list	逗号分隔的好友 ID；可统计社交网络度数
useful	int	其所有评论被标记 “useful” 的总次数
fans	int	关注该用户的人数
average_stars	float	用户给出的平均星级

3.2.4 表间关系

四、数据预处理

4.1 行级过滤与抽样

在最初的行级筛选环节，我们首先依据 city 字段执行精确匹配，仅保留 Philadelphia 所在记录，以消除跨城市经济与文化差异带来的混杂效应。过滤完成后，样本规模从原始的 150 346 家商户缩减至 11 077 家餐饮商户。随后，对

categories、attributes 与 hours 三个半结构化关键字段进行完整性约束：任一字段缺失的行均被删除，以保证后续 JSON 解析与文本处理的成功率

4.2 半结构化字段解析

4.2.1 attributes 字段

针对 attributes 列中存在的双重引号嵌套等异常格式，本文构建了“两级回退”解析机制：先尝试 json.loads，若解析失败再退至 ast.literal_eval；两级均失败则返回空字典。解析成功的嵌套结构通过预先构造的递归函数 flatten_dict 被展开为扁平化列名（如 attr_Restaurants_Delivery），并在展开过程中自动识别并保持 *bool*、*int*、*float* 与 *NA* 类型。最终，借助 convert_dtypes() 将布尔列转换为 Pandas *BooleanDtype*，数值列转换为 *Int64* 或 *Float64*，以便后续高效运算且保留缺失语义。

4.2.2 hours 字段

对于记录商户营业时段的 hours 列，首先将形如 "{'Monday': '8:00-22:00', …}" 的字符串解析为 {weekday : "open-close"} 形式的字典。随后自定义 span_to_hours() 函数处理跨午夜与 24 h 营业情形，将多时段字符串转换为日营业分钟数，再累计得到周营业总时长 weekly_hours。该连续变量在费城子集中分布均值为 69.8 h/周、标准差为 32.0 h。原始 hours 列在生成新特征后即被删除，以降低数据稀疏度。

4.3 缺失值与冗余字段处理

在字段层面，依据缺失率与业务相关性剔除了 26 个冗余列，包括 is_open、postal_code、address 及若干极低覆盖率的属性列，防止模型维度膨胀。解析 attributes 后得到的 12 个布尔列（如 attr_BusinessAcceptsCreditCards 和 attr_Caters）将 <NA> 统一填充为 False，避免在后面的 One-Hot 编码环节引入额外缺失类别。价格区间变量 attr_RestaurantsPriceRange2 被转换为浮点型，并以样本中位数 2.0 进行缺失插补，以稳定数值分布并保留成本层级信息。

4.4 类别与数值特征编码

对四个文本型类别列 attr_WiFi、attr_Alcohol、attr_RestaurantsAttire 与 attr_NoiseLevel，先使用正则表达式对格式进行规范化，再调用 OneHotEncoder 完成 One-Hot 编码，共生成 23 个指示列。除上述四列外的 12 个布尔属性保持 0/1 形式直接输入模型；价格区间列保留为数值特征。至此，餐饮属性侧累计得到 36 维结构化特征。

4.5 文本特征抽取

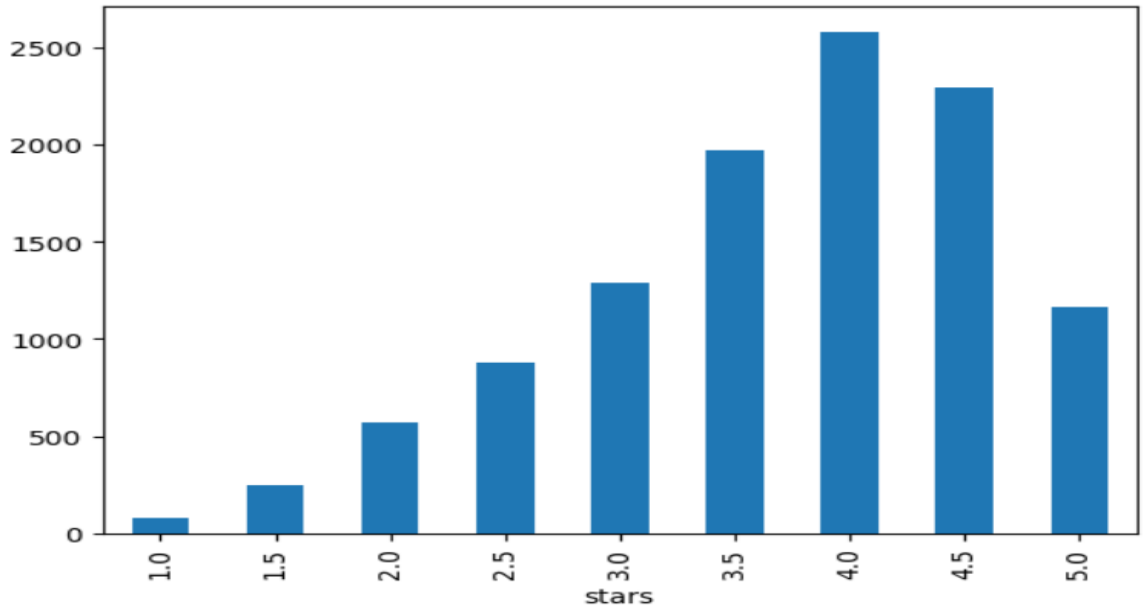
在文本处理阶段，首先对 categories 字段执行全小写化、标点清除及 nltk.word_tokenize 分词操作，并依据扩展停用词表去除高频功能词；随后依次进行 Porter 词干化与 WordNet 词形还原，得到清洗后的属性。为捕获高阶语义，使用 TfidfVectorizer(min_df = 2, ngram_range = (1, 2)) 生成稀疏 TF-IDF 向量，再以 TruncatedSVD(n_components = 100, random_state = 42) 实现潜语义分析（Latent Semantic Analysis），将高维稀疏表示压缩为 100 维密集主题向量 cat_svd_*。原始文本列随后被移除，以避免信息冗余。

经上述步骤处理后，费城餐饮子集最终包含：基础地理与评分信息 6 维、结构化属性特征 36 维、主题向量 100 维，总计 **143 维**；其中布尔列无缺失，数值列缺失率低于 0.1 %，数据类型清晰且缺失可控，为后续评分预测与情感关联建模奠定了坚实基础。

五、数据建模分析

5.1 时空聚类

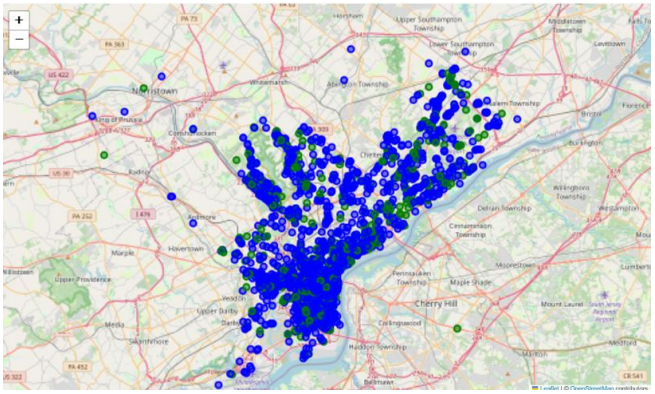
图3：商户星级分布直方图



从商户星级直方图可看出，大部分商户的星级集中在 3.5-4.5 星之间，4.0 星是最常见的评分，1.0 和 5.0 星的商户数量较少。整体来说商户的星级分布偏向较高的评分，意味着大部分顾客对商户的评价偏向好评。

图4：星级<=3.5 商户分布图

图5：星级>3.5 分布直方图(星级=5 标绿)



从不同星级商户的分布图可看出,城市商户主要沿城市主河干及附近交通线路分布，中低星级商户(<=3.5)和中高星级(>3.5)分布区位存在大部分重合，只有少数中高星级商户延伸至城郊方向。因此，仅考虑地理位置的时空聚类并不能挖掘出更多信息。在下面的时空聚类中，我们不仅考虑了商户的地理位置，也将考虑其类别信息。

在本项目中，我们将商户的“类别信息“与”地理位置“结合起来，进行联合的”时空聚类分析“。首先，对商户的 `category_preprocessed` 文本字段进行处理，使用 TF-IDF 向量化方法将其转化为反映类别关键词重要性的高维稀疏向量。随后，为降低维度并提取潜在类别语义结构，我们使用 Truncated SVD 对 TF-IDF 向量进行压缩，得到低维“类别主题”表示。与此同时，对商户的经纬度信息使用 `StandardScaler` 进行标准化，以确保在后续建模中不同特征尺度一致。最终，我们将标准化后的地理坐标与类别主题向量拼接，形成融合商户“类别语义”与“空间位置”的复合特征向量，并在此基础上采用聚类算法（KMeans）识别出同时在**位置接近、业态相似**的商户簇。这一方法有助于发现城市中的特色业态聚集区，为商业选址优化、城市规划以及后续推荐系统提供有价值的数据基础。

图6：商户空间聚类分布图（由肘部法确定聚类中心数量 K = 5,得到聚类结果）

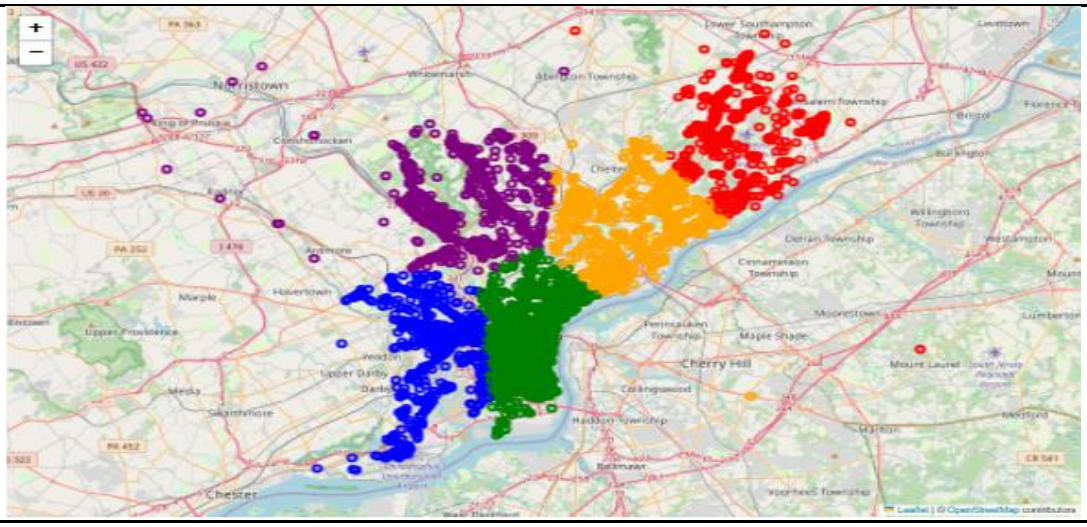


图7：商户聚类分布柱状图

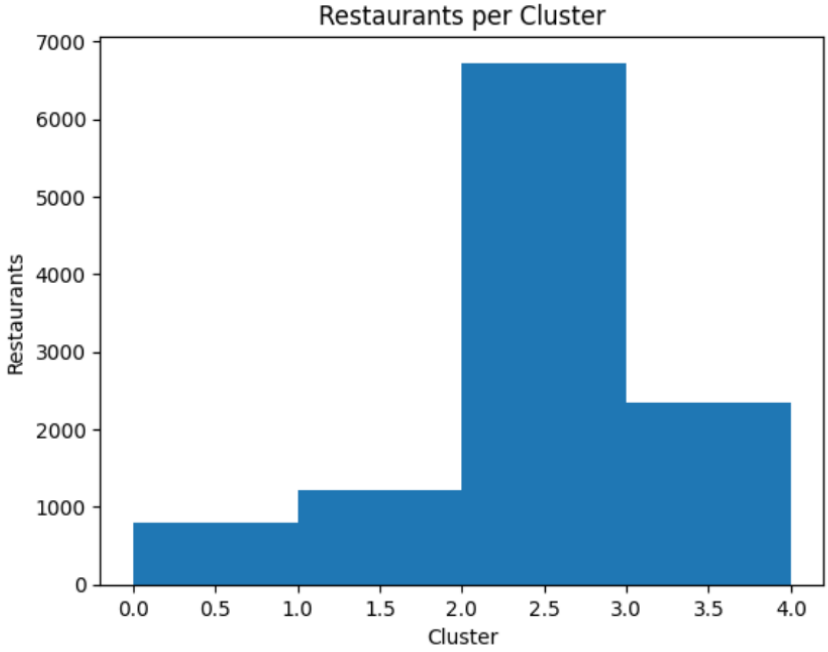


图8：聚类中心 TOP5 类别分布柱状图

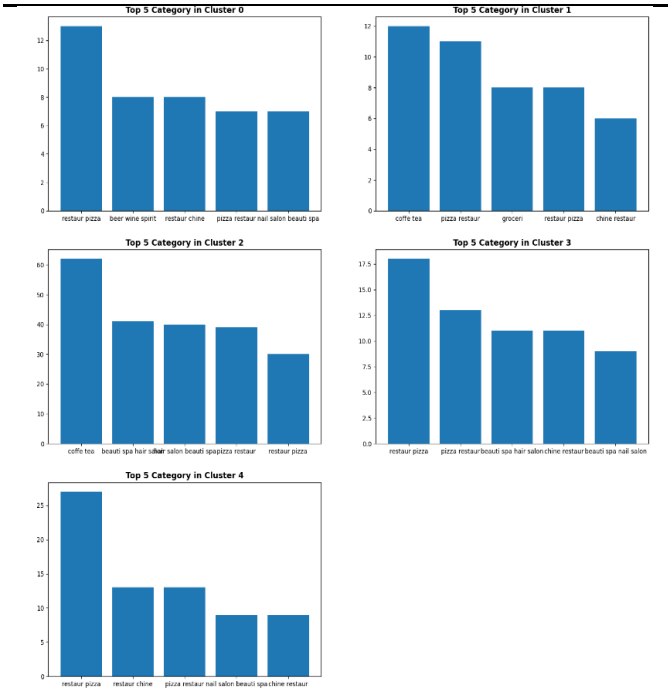
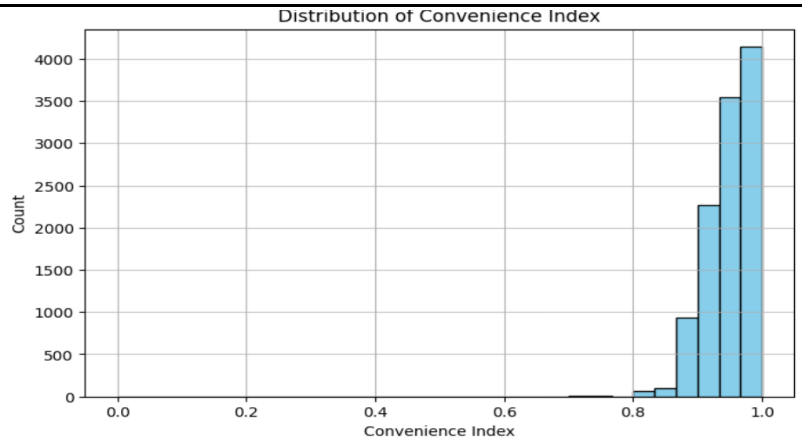


图9：聚类中心类别分布词云图

在本研究中，为了将商户的空间区位优势纳入评分预测模型，我们构造了一个反映地理便利性的代理指标。具体而言，在完成类别与位置联合特征的时空聚类后，我们为每个商户计算其与所属聚类中心之间的欧氏距离，记为 d_i，以衡量其在该类别-空间簇中的“核心程度”。距离越小，说明商户越接近同类商户的地理集中区域，具有更高的潜在客流和地段优势。考虑到原始距离分布的尺度不一致，我们对 d_i进行 Min-Max 标准化处理$d_i^{\text{norm}} = \frac{d_i - \min(d)}{\max(d) - \min(d)}$，并通过反向变换构造出便利性得分 $C_i = 1 - d_i^{\text{norm}}$，使其范围归一化到 [0,1] 区间。该指标越接近 1，表示商户越接近聚类中心、地理区位越优越，因而具备更高的便利性。最终，便利性得分作为一个关键特征，与商户的其他属性共同输入 XGBoost 模型中，用于预测其星级评分，从而增强模型对“地段效应”的刻画能力，提高预测结果的准确性与现实解释力。

图10：商户“便利性”指标直方图



5.2 预测模型 Xgboost

在本项目中，我们构建了一个基于 XGBoost 的回归模型，用于预测商户的星级评分，特别是针对新开业、尚未获得用户评论的商户。为了提升模型的解释性与泛化能力，我们融合了商户的结构化属性信息、文本类目表示以及地理区位优势等多源特征，作为输入变量。

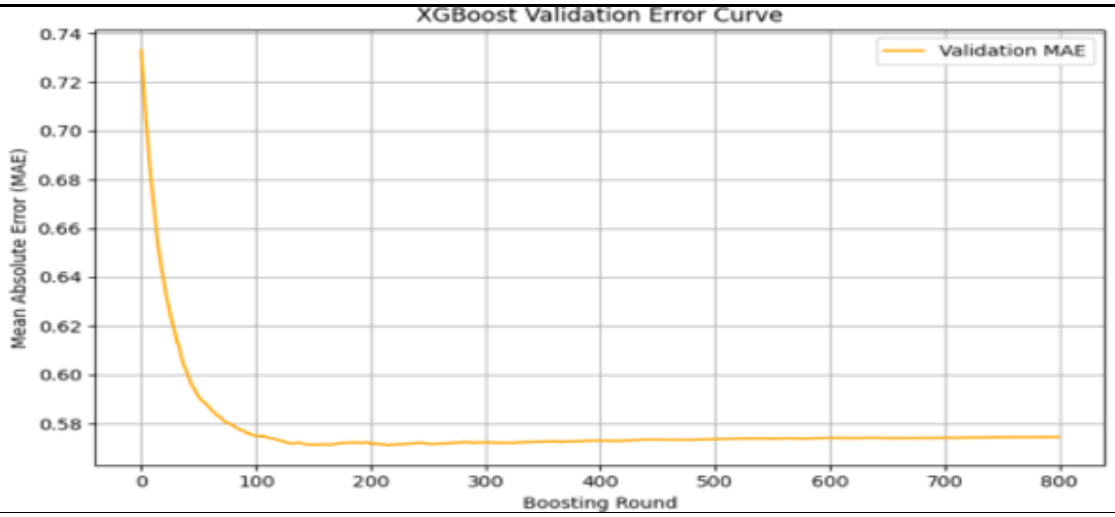
首先，在特征工程阶段构造的字段外，我们对商户名称字段同样进行了 TF-IDF 处理与 SVD 降维，作为补充的语义特征。随后，我们基于类别-空间聚类结果计算每个商户到聚类中心的距离，并采用 Min-Max 反向标准化生成“便利性指数”（convenience index），用于刻画商户所处地段的区位优势。最终，我们将所有处理后的结构化和非结构化特征进行拼接，构建模型输入特征矩阵 X，以商户星级评分 stars 作为预测目标变量 Y。

在建模阶段，我们采用了梯度提升决策树模型 XGBoost，并配置如下参数以提升模型性能：基础树数 n_estimators=800，树深度 max_depth=8，学习率 learning_rate=0.05，子采样比率 subsample=0.8，列采样比率 colsample_bytree=0.8，正则化参数 reg_lambda=1.0，目标函数为平方误差损失 reg:squarederror，评估指标为平均绝对误差（MAE）。训练过程中通过 train_test_split 划分出 80% 训练集与 20% 验证

集，并每 50 轮输出一次验证误差以监控模型收敛情况。

该模型不仅具备处理高维稀疏文本特征的能力，还能整合非线性关系、处理缺失值与多源异构数据，在预测未被评分的新商户时具有较强的实用性与推广价值。

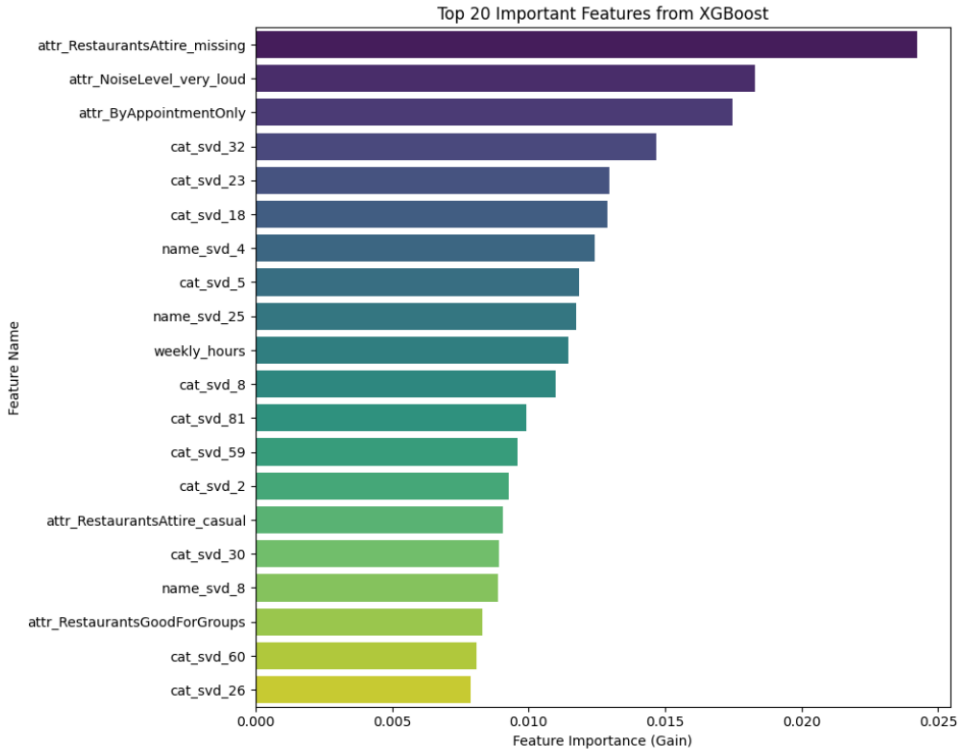
图11：XGBoost 验证误差曲线



从图中可以看到，误差在开始时迅速下降，表明模型在初期的学习过程中对训练数据有较大的改进。随着 Boosting 轮次的增加，误差趋于平稳，说明模型在迭代到一定程度后，已经接近最优，进一步增加轮次并未带来显著的误差降低。

最终，MAE 值稳定在 0.58 左右，说明在最后阶段，模型的预测误差已经趋于稳定。这个误差大小意味着模型的预测结果与真实值之间的平均绝对差距是 0.58。我们的预测目标评分星级是[1,5]步长为 0.5 的区间，该误差水平已处于可接受的范围。

图12：XGBoost 回归贡献 TOP20 字段



由图可见，对 XGboost 预测贡献最大的三个字段分别为 attr_RestaurantsAttire_missing; attr_NoiseLevel_very_loud; attr_NoiseLevel_very_loud。

attr_RestaurantsAttire_missing

意思：该字段指的是商户的穿着要求（Attire）是否缺失。

- 常见值：布尔值（True 表示该商户未填写“穿着要求”信息）。
- 原字段（attire）可能取值：
- "casual"：休闲
- "dressy"：较正式
- "formal"：正式

解释：如果为 True，说明该商户没有在 Yelp 上标注客户是否需要特定着装风格。

attr_NoiseLevel_very_loud

意思：表示该商户的环境噪音水平是否为“非常嘈杂”。

- 常见值：布尔值（True 表示该商户被标记为“very loud”）。
- 原字段（NoiseLevel）可能取值：
- "quiet"：安静
- "average"：正常
- "loud"：较吵
- "very_loud"：非常吵

解释：这通常用于评价餐厅或酒吧的环境，反映出是否适合安静用餐或交谈。

attr_ByAppointmentOnly

意思：该商户是否只接受预约服务。

■ 取值：True 或 False

适用场景：多用于理发店、美容、美甲、诊所等需要预约的服务业。

解释：若为 True，用户必须提前预约才能接受服务；若为 False，可直接上门。

在本项目分析中，针对 Yelp 商户数据中的部分关键属性，我们提出以下优化建议，以助力商户在缺乏用户评论的情况下，通过改善自身服务和展示方式，提升星级预测评分。首先，针对 `attr_RestaurantsAttire_missing` 字段，商户应主动完善平台上的属性信息，如着装要求、付款方式等，避免信息缺失对模型预测产生负面影响，同时提升用户对商户专业度的感知。其次，对于 `attr_NoiseLevel_very_loud` 所反映的用餐环境问题，建议商户通过降低背景音乐音量、设置隔音区域等方式优化噪音管理，改善整体用户体验。最后，`attr_ByAppointmentOnly` 表示仅限预约服务的限制性机制，可能在一定程度上降低用户的到访便利性。商户可在保障运营秩序的同时，提供部分时段的自由到访服务，提升灵活性和用户粘性。上述措施不仅有助于优化顾客实际体验，也有利于模型在多源特征输入下做出更高评分的预测，助力商户在智能推荐和平台排序中占据更优位置。

5.3 评论文本情感分析

在评论文本情感分析模块中，我们旨在通过自然语言处理技术提取用户评论中的情绪倾向，为商户评价提供更具解释性的辅助指标。首先，我们以验证集中的商户 `business_id` 为主键，与 Yelp 的 `reviews` 文件进行左连接，提取每个商户最多前 50 条评论，模拟试运营期间的用户反馈。随后对评论文本进行标准化预处理：**词条化（Tokenization）** 将文本拆分为基本分析单元（单词或短语）；**规范化（Normalization）** 操作如统一大小写、简繁转换等，消除格式差异；**噪音去除（Noise Removal）** 包括剔除 HTML 标签、特殊字符及停用词（如“的”、“是”等），提升语义纯度。

在情感分析环节，我们利用情感分析工具（如 `TextBlob`）对每条评论打分，生成极性分数（polarity），范围为 $[-1, 1]$。设定阈值 polarity > 0.2 为“好评”，将每个商户的好评数量与评论总数计算得出“**好评率**”指标，反映该商户在试运营期间的用户满意度水平。需要指出的是，该“好评率”**并未直接作为 XGBoost 模型的输入特征**，而是作为模型预测结果的补充评估维度。最终，我们根据商户的星级预测值和好评率进行双重排序，依次降序筛选出前 10 个表现优秀的新开业商户，为平台推荐机制与商户评级提供数据驱动支持。

图13: 最后得出的前 10 优秀商户

	business_id	stars	positive_rate
78	6085NRg7QH3vXpc50F4UHQ	5.0	1.000000
312	QJZdu9kFpKh4Fy8_YuvvXQ	5.0	1.000000
423	YbpNzwI5iBvsBDwdwc9Mmg	5.0	1.000000
652	tyFuhfn1BDGHWpmwpkwBAQ	5.0	1.000000
284	NDHgJsy-4Lb6WhERPXPs0A	5.0	0.928571
258	Ktg3ahlxk0JlkJwXAqu2ew	5.0	0.920000
472	cVV8GWVIe9BwyCOKwrFgPA	5.0	0.860000
510	fq1yCVBgBB7s6V-D68NO1g	5.0	0.837209
79	609Lr-Hvo3sr9amdiimOJA	5.0	0.818182
263	LHzg5i6hX1Qb3OxtWJ4QDQ	5.0	0.800000

5.4 社交网络构建与推荐实施

5.4.1 整体介绍

在 Yelp 数据集中，用户之间存在显式的“朋友关系”，为基于社交网络的个性化推荐提供了数据基础。本项目以验证集中综合“星级预测”和“好评率”指标选出的 TOP10 优质新商户为目标推荐对象，进一步挖掘用户的社交图谱信息，实现“好店优先推荐给朋友”的传播策略。[10]

首先，我们通过解析 user_id 和 friends 字段构建了社交网络图，其中节点为用户，边为好友关系，默认边权重为 1。共提取出约 36 万条有效的用户间关系。为控制计算资源消耗并提升推荐的针对性，我们采用滚雪球抽样[9]方法，从度中心性最高的节点出发，限定最大深度为 5、最大节点数为 10,000，从而获取网络的关键结构子图。

随后，我们计算用户的图结构指标（包括度中心性与介数中心性），识别出若干“关键传播用户”作为种子用户。通过优先向种子用户推送优质商户信息，利用其在网络中的传播优势，实现了更高效的推荐扩散。

最后，为提升推荐系统的个性化水平，我们对社交网络进行社群划分，应用 Louvain 算法识别网络中的潜在社群结构。结合每个社群的模块度与连接模式，我们制定面向不同社群的推荐策略，实现“用户群体细分—内容定制”的推荐闭环。
该模块将社交网络结构、用户活跃度与信息扩散机制引入商户推荐任务中，不仅增强了模型的可扩展性与传播效率，也提高了推荐的用户相关性与接受度。

5.4.2 提取边、节点

社交网络中一共有 3 个要素：节点、边以及边权重。首先，我们根据 Yelp 用户相关数据中“user_id”字段构建节点，并过滤掉重复值以及不符合用户 ID 范式的异常值。其次，按分号分割“friends”字段，还可解析出所有用户间的关系对（source-target），共计 36 万对。对于边权重，鉴于 Yelp 只提供了用户间是否为朋友的信息，缺少表现用户间关系亲疏远近的有关数据，我们不妨假定所有边权重都为 1。

5.4.3 滚雪球抽样

根据上述分析，不难发现全样本的社交关系网络较庞大，直接运用全样本数据进行推荐可能面临以下缺点：

1）处理大规模社交网络数据需要耗费大量的计算资源和存储空间，难以在合理时间内完成模型训练和推荐生成。

2）全样本数据中可能包含大量与目标推荐无关的用户和关系，这些冗余信息会干扰模型的学习过程，降低推荐的准确性和相关性。

图14：全样本数据的社交网络图

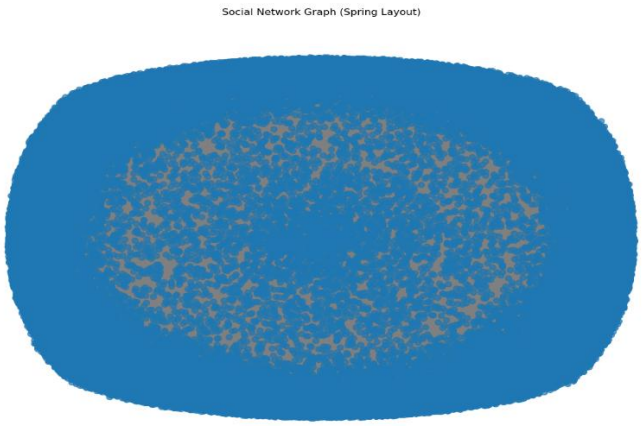
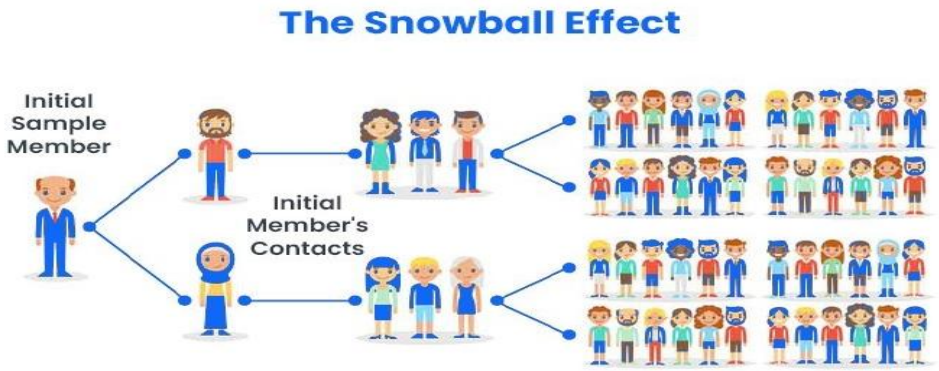


图15：滚雪球效应图示



为解决上述问题，我们采用滚雪球抽样方法，从一个或多个初始节点开始，逐步扩展到其邻居节点，从而快速获取网络的局部结构和特征。具体步骤如下：

首先，我们选择度中心性最高的节点作为初始节点。度中心性越高，该节点在网络中的连接性越强，能够快速将信息传播到更多的节点，符合我们基于社交网络关系进行优质新商家推荐的目标。相比之下，随机选取节点作为起始节点的抽样方式不是最

优解，因为随机选择的节点的连接性较弱，无法有效覆盖网络的关键部分。选择度中心性高的节点作为起点，可以确保在有限的抽样范围内，获取到更具代表性和价值的网络结构信息，从而提高推荐系统的准确性和效率。

接下来，我们需要设定最大节点数（max_node）和最大深度（max_depth）来控制抽样范围。具体而言，我们选取最大节点数为 10000，最大深度为 5。通过限制最大节点数和深度，我们能够避免因样本过大而导致的计算资源浪费，同时确保抽样结果能够涵盖网络的关键部分，为后续的推荐算法提供足够的信息支持。

最后，我们从初始节点开始，逐步访问其邻居节点，并将这些节点加入样本集。每一步根据设定的最大节点数和深度限制，动态调整样本集的规模。

图16：多个随机节点为起点的滚雪球抽样

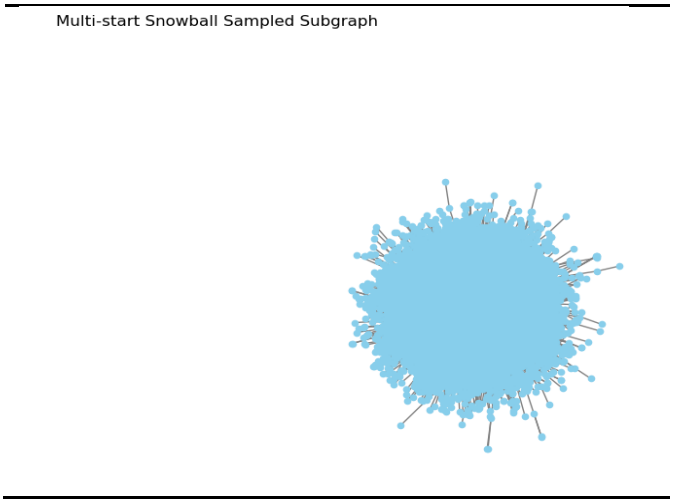
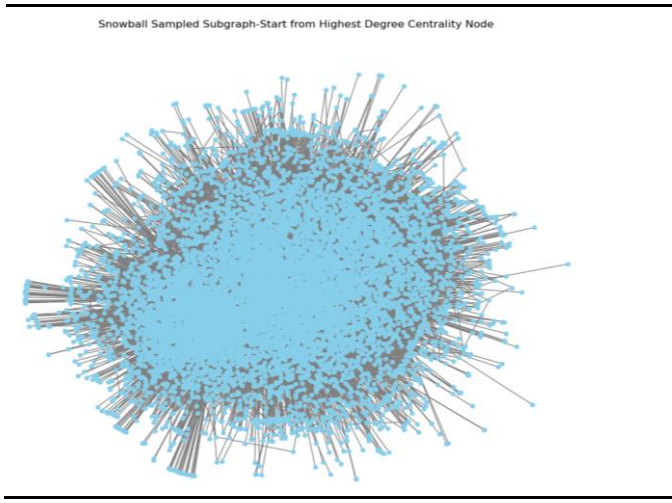


图17：度中心性最高点为起点的滚雪球抽样



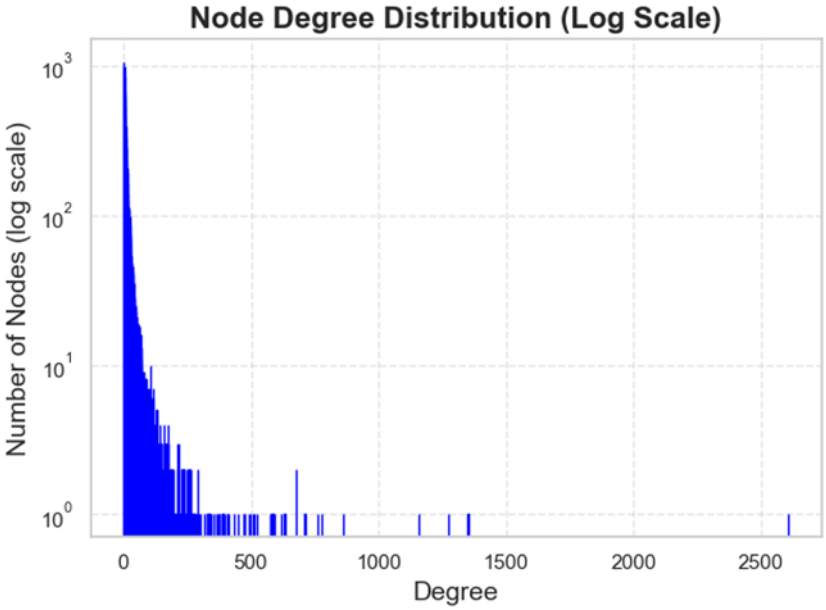
5.4.3 优质用户识别

为确保推荐系统的有效传播和高影响力，我们进一步识别社交网络中的种子用户。种子用户在网络传播中起着关键作用，能够快速将信息传播到更广泛的用户群体。识别种子用户的步骤如下：

首先，我们计算每个用户的多种社交图指标，包括度中心性（Degree Centrality）和介数中心性（Betweenness Centrality）。度中心性衡量节点的连接数量，反映了用户在网络中的活跃度和连接性；介数中心性衡量节点对信息传播路径的控制力，反映了用户在网络中的中介作用。这些图指标帮助我们识别“关键传播用户”。

我们根据计算得到的指标，筛选出具有高度中心性和高介数中心性的用户作为种子用户。在推荐新商家时，优先将信息推送给这些种子用户，利用他们的高传播能力，快速将推荐信息扩散到整个社交网络。通过这种方式，我们能够提高推荐信息的传播效率和覆盖范围，增强推荐系统的影响力和效果。

图18：抽样后节点的度中心性分布



5.4.4 社群识别

在完成滚雪球抽样并构建社交关系网络后，我们进一步对网络进行社群识别，为个性化推荐提供更细致的用户群体划分。社群识别的步骤如下：

首先，我们对抽样后的社交网络应用社群检测算法。基于网络具有拓扑结构，运用 Louvain 算法我们能够在大规模网络中高效地发现社区结构，即将具有相似连接模式的节点划分到同一社群中。Louvain 算法主要包括两个阶段：首先将每个节点视为一个独立社区，逐步调整其归属，使模块度最大化；然后将形成的社区折叠为新节点，重复上述过程，直至模块度变化量低于阈值。

最后，我们将识别出的社群信息整合到推荐系统中。根据用户的社群归属，为每个社群定制个性化的推荐策略。例如，对于具有相似兴趣和社交关系的用户群体，推荐与该社群特征相关的优质新商家。通过社群识别，我们能够更精准地捕捉用户的社交偏好和行为模式，从而提高推荐的相关性和多样性，增强用户对推荐结果的满意度和接受度。

图19：基于 Louvain 算法的社群识别

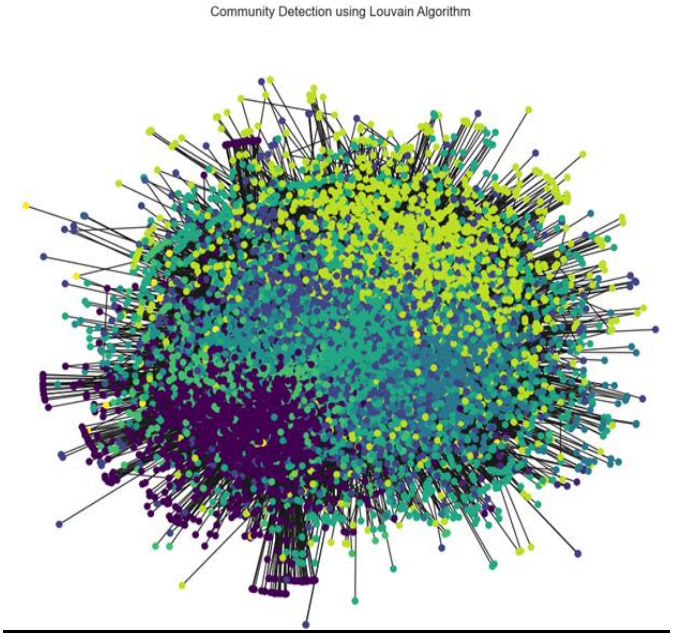


图20：最优社群划分结构

六、数据分析结论总结

6.1 基于多源结构化特征的 XGBoost 商户星级预测模型构建

本研究构建了一个集成学习回归模型，采用 XGBoost 算法对 Yelp 商户的最终星级评分进行精准预测。特征设计阶段，我们系统性整合了商户的类别信息（经 TF-IDF 向量化与 Truncated SVD 主题压缩处理）、地理空间位置（经 StandardScaler 标准化处理）、服务属性（如是否接受信用卡、是否适合儿童等二值化标签），并引入空间聚类计算得到的“便利性”指标，作为衡量商户与城市功能中心距离的反向得分。通过上述多源结构化特征的融合与建模，我们有效提升了模型对新开业商户（无评论数据）的星级预测能力，尤其解决了平台冷启动场景下的评价空缺问题，为智能化商户评估与准入推荐提供了数据驱动的技术支持。模型验证阶段以 MAE（Mean Absolute Error）为评估指标，获得了令人满意的误差控制，验证了模型的实用性与泛化能力。

6.2 融合主观情感极性与客观评分的多维优秀商户识别机制

在完成基于客观属性的星级预测之后，项目进一步引入用户评论数据，以主观体验维度刻画商户服务质量。我们对验证集中的商户 ID 与原始评论数据进行关联，并基于评论文本进行分词、规范化、噪声清洗及停用词移除等预处理流程。随后，采用情感分析工具对每条评论计算极性得分（Polarity），并设定合理阈值（>0.2 视为好评）构建“好评率”指标。该指标反映了试运营阶段用户的主观满意程度。我们基于商户的星级预测结果与好评率联合排序，筛选出表现稳定、用户反馈积极的 Top10 优质商户。这种主客观融合的评价机制不仅提升了商户甄别的精度，也为平台后续运营资源配置（如流量倾斜、推荐位分配）提供了更具解释力和透明度的量化标准。

6.3 挖掘用户社交图谱实现基于网络结构的个性化传播推荐策略

考虑到社交网络在信息传播与影响力扩散中的关键作用，我们进一步利用 Yelp 用户数据构建了社交图谱。具体做法是解析用户之间的好友关系，构建无向边权网络，并采用滚雪球抽样策略，以度中心性最高的用户为起点，在设定最大节点数与深度（如 max_node = 10000, max_depth = 5）的约束下，抽取子图用于传播建模。在该网络基础上，我们识别出一批“种子用户”，即在社交结构中拥有高介数中心性和高连接度的节点，优先将优质新商户信息推送至这些用户的邻接网络，最大化信息扩散效果。此外，我们引入 Louvain 社群检测算法，对抽样网络进行社群划分，实现了面向群体的定向推荐策略。每个社群代表具有相似兴趣和社交行为特征的用户集合，可进一步定制化推荐方案。该策略不仅提升了推荐结果的相关性和多样性，还显著增强了平台用户粘性和口碑传播能力，展示了社交驱动推荐在本地生活服务场景中的强大潜力。

七、项目实践价值及展望

7.1 项目价值

综上所述，项目最终构建了一个集成时空特征挖掘、文本情感理解与社交图谱传播机制的智能推荐系统，专为本地生活服务平台中“新商户”的早期识别与用户匹配设计。在冷启动问题中，系统首先依托 XGBoost 模型对商户结构化信息（如类别、空间位置、便利性）进行精准评分预测，弥补新商户缺乏历史评论的困境；随后引入试运营期间的用户评论，通过自然语言处理和情感极性打分，建立“好评率”这一用户满意度代理指标，增强星级预测的可解释性与多维筛选能力；在此基础上，系统进一步构建用户社交网络，运用图计算方法实现种子用户定位与社群划分，最终以“好店优先推荐给朋友”的策略，在社交传播层面推动优质商户的精准触达与扩散。该推荐系统不仅融合了地理位置的便利性分析、用户评价的情感智能识别以及人际关系网络的传播潜力，也展现出良好的可扩展性和实用价值，为城市餐饮平台在新商户孵化、用户冷启动推荐、精准营销等方面提供了数据驱动的整体解决方案。

7.2 项目缺陷

尽管本项目构建了较为完整的商户星级预测与推荐系统，但在模型效果和特征设计方面仍存在一定不足，后续工作可从以下几个方向进一步优化：

1. 模型预测性能仍有提升空间

在当前的数据集基础上，我们构建的 XGBoost 回归模型在验证集上的平均绝对误差（MAE）为 0.57，表明模型在拟合商户星级方面仍存在一定误差。造成该问题的可能因素包括：（1）训练样本数量有限，Yelp 在费城区域记录的有效商户信息相对较少，难以全面捕捉特征分布；（2）商户的结构化属性信息较为稀疏，模型能够学习的有用信号受限；（3）文本类特征（如类别、商户名等）经 One-Hot 或 TF-IDF 编码后维度较高、信息密度低，噪声影响显著。

为改善模型性能，未来工作可尝试：扩大样本规模，引入更大城市（如纽约、洛杉矶）的多源数据；提升特征工程深度，例如引入商户历史评分变化趋势、类别层级信息、用户画像等；并结合网格搜索或贝叶斯优化等方法，系统性地调优 XGBoost 参数，如 max_depth、learning_rate、subsample 等，从而提高模型的泛化能力。

2. 关键特征“便利性指标”在模型中的表现有限

本项目引入了结合空间位置与类别信息的“便利性”指标，即每个商户到最近聚类中心的归一化反向距离，旨在刻画地理区位优势。然而，XGBoost 特征重要性分析显示该指标在整体特征中的排名靠后，未能充分提升模型解释力。这可能源于以下几点：（1）模型已通过其他强特征（如商户评分历史、类别主题）捕捉了大部分预测信号；（2）便利性指标与星级评分之间的相关性较弱，未能准确反映用户体验；（3）KMeans 聚类对聚类数 k 的敏感性可能导致中心划分粗糙，影响指标分布的区分度。

对此，后续研究可尝试采用更灵活的非参数聚类方法，如 DBSCAN、HDBSCAN，识别真实的地理商圈结构；同时在便利性度量中引入加权中心或多中心模型，更精细地反映地理可达性。此外，可结合地图 API 或道路网络数据，构建基于交通路径的实际可达性指标，以增强“便利性”的解释力与现实意义。

3. 便利性指标可进一步结合人流密度与交通可达性优化

目前构造的便利性指标主要基于商户与空间类别聚类中心的欧氏距离进行度量，虽然一定程度上体现了商户的空间集聚程度，但忽略了真实的人流分布与交通可达性等关键影响因素，限制了指标对实际消费便利性的刻画能力。

为提升便利性指标的现实性与预测力，未来可引入多源城市空间数据（如 POI 热力图、地铁公交站点、道路网络、街区步行性指标等），将静态地理位置信息扩展为动态人群行为的代理变量。例如，可利用高德或 Google Maps API 计算商户与最近交通枢纽间的通达时间（而非几何距离）；或使用手机信令、人流轨迹等数据构建“客流权重图”，量化商户实际被访问的潜力。

此外，还可借助 GIS 工具建立可达性模型，基于交通网络分析不同时间段的“到达时间等值线”，从而构造更贴近现实消费路径的便利性评分。这类融合式的便利性设计，有望提升特征对消费者行为的解释力，在商户星级预测和潜力识别中发挥更关键作用。

4. 文本挖掘中情感分析阈值设置存在主观性，分类性能有待验证

在试运营期间商户优劣评估过程中，本文采用情感分析工具对评论文本进行打分，并将极性得分（polarity）大于 0.2 的评论视为“好评”，据此计算“好评率”指标。然而，这一阈值的设置具有一定的主观性，缺乏与真实人工标注情绪标签的对比验证，导致情感分类的准确性无法量化评估。

在无监督语义评分场景下，阈值的选择实质上是一个超参数（hyperparameter），对结果有显著影响。设定过高可能低估正向情感密度，过低则可能误判中性或负向表达为正向。此外，情感极性在不同城市文化、评论风格和领域词典中可能呈现不同分布，因此单一静态阈值未必具备普适性。

为改进该问题，未来可采用以下优化路径：

- **构建小规模标注集：**人工标注一部分评论的情感倾向，用以评估当前阈值下的 Precision / Recall，并探索更优分界点；
- **引入监督学习方法：**利用标注数据训练情感分类器（如 Fine-tuned BERT、SVM

等），摆脱阈值依赖；

- **分布自适应阈值设定：**采用 Otsu 分割、聚类等方式在情感得分的实际分布中自动划定阈值，提高适应性与客观性。

通过更科学合理的情感分类策略，有望显著提升“好评率”作为主观评价指标的有效性与可信度，从而增强推荐结果的可解释性和决策参考价值。

5. 社交网络推荐存在 B 端与 C 端断连问题，影响推荐覆盖能力

在构建基于社交网络的商户推荐系统时，我们采用滚雪球抽样从用户网络中提取局部结构，并基于用户的朋友关系向其推荐验证集中筛选出的 Top 优质商家。然而，在实际应用中发现，优质商家（business_id）与社交网络中抽样得到的用户（user_id）之间存在结构性“断连”，即部分起始节点及其邻居用户从未在试运营期间对这些商户发表过评价。此类断连现象导致推荐路径无法有效打通，从而削弱了“好店推荐给朋友”的策略效果。**其产生原因主要包括：**

- **评论行为稀疏：**多数用户评论数量有限，导致试运营期间参与评论的用户与雪球样本重合度较低；
- **起始节点选择偏差：**即便具备高中心性的节点，其社交影响力并不一定覆盖对 Top 商户感兴趣的目标用户；
- **地域分布不一致：**部分商户与抽样用户可能位于不同地理簇，社交网络结构与地理兴趣点之间存在偏移。

为改善该问题，后续可从以下方向展开优化：

- **多中心滚雪球抽样：**以多个评论过 Top 商户的用户为起点，提升 B 端与 C 端的连接概率；
- **跨模态用户扩展：**结合评论相似性、地理邻近性等非结构性信息对社交图进行“软扩展”；
- **引入潜在兴趣传播模型：**如基于图卷积（GCN）或 LightGCN 等模型，在弱连接结构中挖掘间接兴趣传播路径。

通过优化网络采样与连接机制，可有效增强商户与潜在用户之间的社交联系，提升推荐系统的可达性与传播效果。

7.3 项目展望

1. 融合推荐算法，实现社交与兴趣的双重驱动推荐

当前系统主要依赖于用户之间的社交关系实现“朋友推荐朋友”的传播机制，推荐策略在一定程度上受限于社交网络结构的连通性。未来可引入协同过滤、矩阵分解（如 SVD++）、基于图的推荐（如 LightGCN）等主流推荐算法[16]，挖掘用户历史行为（如浏览、点击、收藏、评论）背后的兴趣偏好。通过社交推荐与内容推荐相结合，不仅可以实现“你朋友喜欢什么你也可能喜欢”，还可以根据“你曾喜欢什么”进一步预测潜在兴趣，实现更加个性化、精准且具有多样性的推荐。

2. 提高推荐系统的多样性与覆盖率

目前社交网络在传播过程中易形成“信息回音室”，即用户仅接收到自己朋友圈或社群内的推荐信息，可能遗漏更广泛的优质商户。引入推荐系统算法后，可基于用户兴趣画像与潜在偏好生成补充推荐项，发掘用户尚未意识到但可能感兴趣的内容，从而增强推荐系统的探索性与内容多样性，拓展推荐覆盖范围，打破社交推荐的局限性。

3. 深度挖掘评论文本价值，引入时间序列建模进行趋势预测

目前对用户评论的处理主要基于情感极性评分构建“好评率”指标，缺乏对评论动态变化的建模。未来可引入深度学习模型（如 LSTM、BERT + Transformer 结构）[14][15]，结合评论发布时间，对评论文本序列进行建模与预测，识别商户口碑随时间演变的趋势，从而为商户星级变化趋势建模、潜在危机预警和营销决策提供更具时效性与洞察力的依据。

通过上述扩展与优化，本项目最终可发展成为一个兼顾社交关系传播、个性化兴趣预测、文本动态理解的复合型推荐系统，真正服务于新商户开业初期的用户引流与口碑建设需求。

参考文献

[1] Bhanu Prakash Reddy Guda, Mashrin Srivastava, Deep Karkhanis.Sentiment Analysis: Predicting Yelp Scores[C].Carnegie Mellon University, 2023.

[2] Yang, G., Xu, X., & Zhao, F. (2019). A user rating prediction model based on XGBoost. Data alysis and Knowledge Discovery, 3(1), 118–126.

[3] Jin, T. (2020). Business Data Analytics – Yelp Dataset. arXiv preprint arXiv:2001.xxxx.

[4] Elkouri, A. (2019). A Machine Learning Based Yelp Recommendation System. Stanford 229 Report.

[5] Fan, M., & Khademi, M. (2014). Predicting a Business Star in Yelp from Its Reviews Text one. arXiv preprint arXiv:1401.0864.

[6] Liu, S. (2020). Sentiment Analysis of Yelp Reviews: A Comparison of Techniques and Models. iv:2010.xxxx.

[7] Elkouri, A. (2015). Predicting the Sentiment Polarity and Rating of Yelp Reviews. Stanford 229 Report.

[8] Guda, S., & Others. (2022). Sentiment Analysis: Predicting Yelp Scores Using BERT and ention Mechanisms. OAJ Artificial Intelligence & Machine Learning.

[9] Goodman, L. A. (1961). Snowball Sampling. The Annals of Mathematical Statistics, 32(1), 148–170.

[10] Li, X., Xu, G., Chen, E., et al. (2015). Learning User Preferences across Multiple Aspects for Merchant Recommendation. IEEE ICDM.

[11] Stanford SNAP. (2017). Predicting Yelp Reviews. CS224W Social and Information Network Analysis Final Report.

[12] Crain, B., et al. (2015). An Evaluation of Yelp Dataset Structure. Stanford CS229.

[13] Zhang, Y., & Wang, L. (2023). Social Networks and Consumer Behavior: Evidence from Yelp. ScienceDirect.

[14] Bhatt, C., & Patel, K. (2023). Polarity of Yelp Reviews: A BERT–LSTM Comparative Study. Information, 14(5), 260. https://doi.org/10.3390/info14050260

[15] Liu, Zefang. (2020). *Yelp Review Rating Prediction: Machine Learning and Deep Learning Models*. arXiv preprint arXiv:2012.06690. https://arxiv.org/abs/2012.06690

[16] Zhang, Shuwei, Tang, Maiqi, Zhang, Qingyang, Luo, Yucan, & Zou, Yuhui. (2021). *Given Users Recommendations Based on Reviews on Yelp*. arXiv preprint arXiv:2112.01762. https://arxiv.org/abs/2112.01762