美国 LAS VEGAS 餐饮业商户网络口碑影响因素分析

2021-4-14

姓名：曹恺燕 学校：复旦大学 学号：20210980133
姓名：易经杰 学校：复旦大学 学号：20210980073
姓名：张志远 学校：复旦大学 学号：20210980131

美国 Las Vegas 餐饮业商户网络口碑影响因素分析

1 背景介绍与研究问题

随着互联网的快速发展，消费者能够通过网络等渠道对商家进行评价，同时商家也由此形成自己的网络口碑。网络口碑的好坏对消费者的消费选择和商家的经营发展都会产生重要的影响。因此，研究商户网络口碑的影响因素具有非常强烈的现实意义。本文基于 Yelp 网站在实际运营中产生的消费评论数据，以拉斯维加斯餐饮类型的商户为研究对象，通过数据分析，得到网络口碑的影响因素及其重要程度，并由此为商家提高网络口碑提供实际的建议。

2 数据的说明与描述

2.1 Yelp 数据集

Yelp 是美国最大的点评网站，创立于 2004 年，它囊括各地餐馆、购物中心、酒店、旅游等各个领域的商户，用户可以在 Yelp 网站中给商户评星级（1 到 5 星）、提交评论、交流消费体验等等。本文收集了 2004 年 10 月 12 日至 2019 年 12 月 13 日间的用户、商户和评论信息，共 1968703 名用户、209393 个商户、8021124 条评论。用户信息包括每个用户的注册时间、评论数目、平均星级等；商户信息包括每个商户的地理位置、营业种类、营业时间、环境设施、平均星级等；评论信息包括每条评论的具体星级、所评论用户、所评论商户、评论内容等。

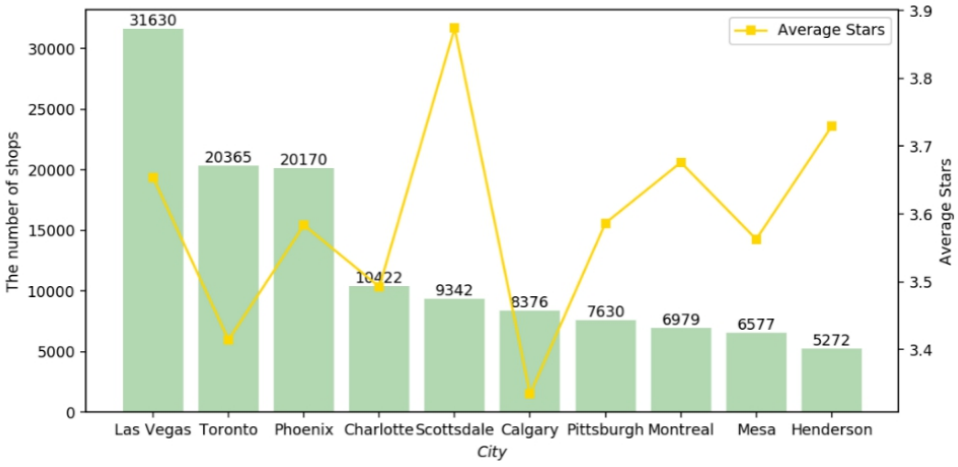


图 1 全美国所在城市商户数（前 10）以及平均星级图

数据集共包括 1251 个城市的商户信息，由图一可知，Las Vegas、 Toronto 和 Phoenix 的商户数量最多，分别为 31630、20365 和 20170。同时可看到在商户数量前 10 的城市中，Scottsdale 商户平均星级最高，Calgary 最低。考虑到样本数目和数据代表性，本文以 Las Vegas 作为研究对象。

2.2 Las Vegas 餐饮业数据集

考虑到数据的时效性，本文筛选出 2010 年 1 月 1 日至 2019 年 12 月 13 日 Las Vegas 商户及其相关评论，并根据商户属性筛选出餐营业商户，最后得到 Las Vegas 餐饮业数据集，包括商户信息和对应的用户评论信息。

表 1 重要数据含义解释

数据类型	英文符号	中文名称及含义	数据类型	英文符号	中文名称及含义
文本类	Business_id	商业 id 号码	数值类	Stars	星级
	Name	商户名称		Review_count	评论数量
	Address	商户地址		Hour_daily	平均每天营业时间
	City	商户所属城市		Neighbor_100(200.etc)	商户周围 100(200)米内其他商户数
	State	商户所属州		Star_b2019	2019 年前的未加权评论平均评价星级
	Postal_code	邮编		Star_b2019_weighted*	2019 年前加权评论平均评价星级
	Attributes	属性		Star_b2019_log_weighted*	2019 年前 log 加权评论平均评价星级
	Categories	类型		Stars_weighted_2019	2019 年的加权评论平均评价星级
	Cate_main	商户细分类别		Stars_log_weighted_2019	2019 年的 log 加权评论平均评价星级
	Latitude	纬度	布尔类	Out_door	是否能户外用餐(是取 1,否取 0)
	Longitude	经度		Credit_card	是否支持信用卡(是取 1,否取 0)
数值类	Week_open	周末开业天数		Wifi	是否提供 Wifi(是取 1,否取 0)
	Days_open	每周开业天数		Group	是否适合多人聚餐(是取 1，否取 0)
	Open_hour_wd	周内平均开业时间		Business_parking	是否有停车场（是取 1，否取 0）
	End_hour_wd	周内平均打烊时间		Reserve	是否支持预定(是取 1,否取 0)
	Open_hour_we	周末平均开业时间	等级类	Price	价格范围(1 到 4,数字越大越贵)
	End_hour_we	周末平均打烊时间		Noise	噪音水平(quite:0,average:1,loud:2)

*注：加权星级指的是以该区间段评论字数为 weight 的商户星级，log 加权评价星级指的是以 log(评论字数 + 1)为 weight 的商户星级（计算过程见式（2））

2.2.1 商户描述性统计

绘制商户平均星级图（以 0.5 为单位取整）和商户评论条数图，如图 2、3 所示：

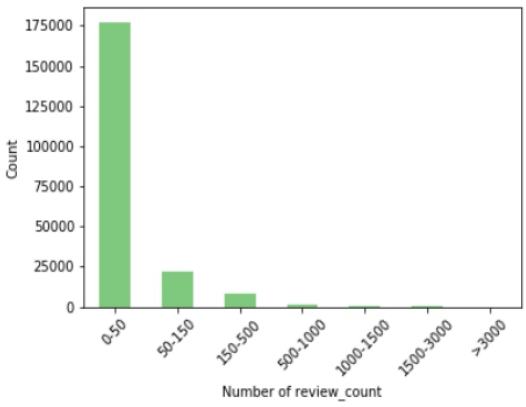


图 2 不同评价星级商户数目

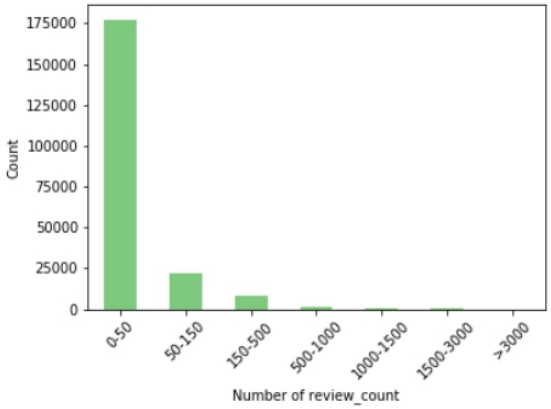


图 3 商户评论数目分布图

由图中可知，大部分商户的评价星级都在[3.5,4.5]区间范围内内，评价星级为 4.0 星的商户最多，共 2043 家。同时，商户评论数量呈幂律分布，多数商户评论少于 50 条，极少数商户评论超过 3000 条。

根据商户平均评价星级，可将商户分成三种类型：差评、中评和好评（划分阈值分别为 2.5 星和 4 星）。则可得到三种类型餐厅的人气状况，如图 4 所示。由图中可知，差评、中评和好评餐厅的评论数量依次增加。

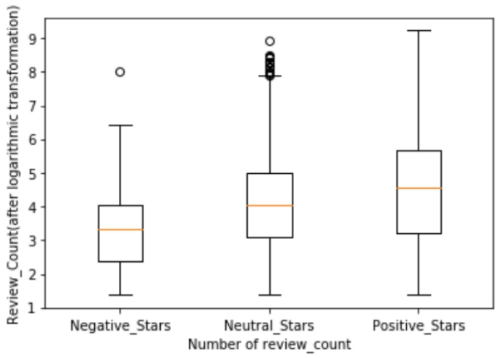
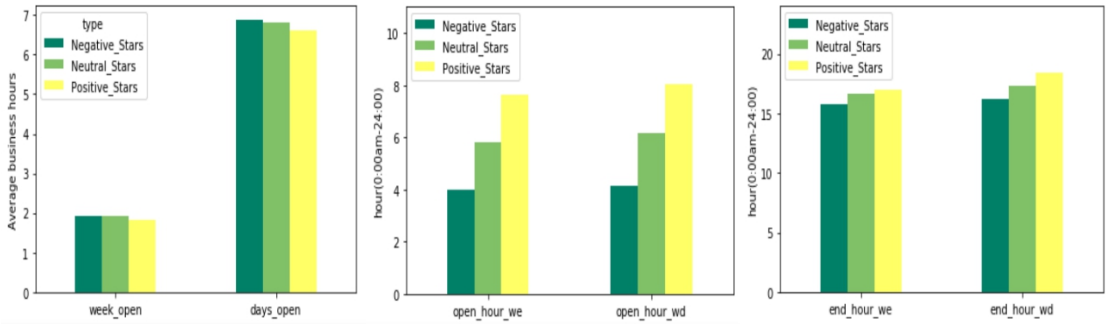


图 4 三种网络口碑类型商户评论数量（取对数）

同时，分别绘制三种类型餐厅周末和周内平均营业时间、开业时间和打烊时间，如图 5(a)、(b)、(c)所示。可以发现，差评、中评和好评餐厅平均营业时间逐次降低，且好评餐厅比前两种餐厅的开业时间和打烊时间都稍晚。



(a) 营业时间 (b) 开业时间 (c) 打烊时间

图 5 三种网络口碑类型商户每日（周末和周内）平均营业时间、开业时间和打烊时间柱状图

结合各餐厅的地理位置信息，绘制三种餐厅的地理位置图及评论信息热力图，如图 6 所示。蓝、黄、红三色分别代表差评、中评、好评餐厅，可观察到在市中心地区餐厅分布较多，三种类型的餐厅都有一定的比例，同时还可以发现在市边缘也存在着好评商户。结合图 7，发现市中心地区餐厅的评论数目较多，即人气较旺。

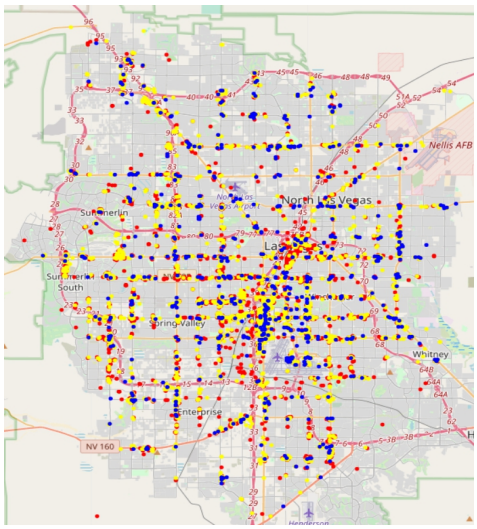


图 6 三种网络口碑类型餐厅地理分布情况

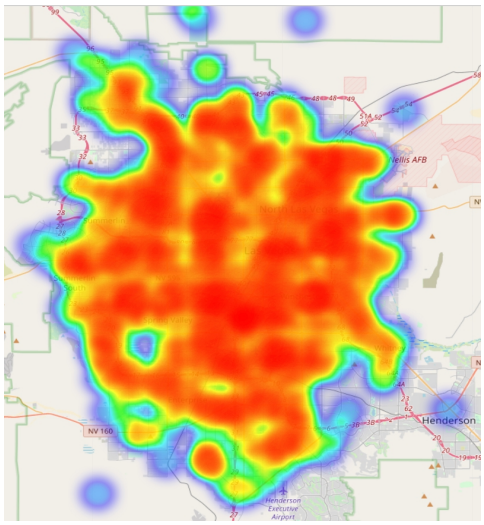


图 7 三种网络口碑类型餐厅评论数目热力图

2.2.2 用户评论描述性统计

据处理得到的用户评论信息数据集，可绘制出用户评论总数与平均评价星级图，如图 8 所示。总体来看，2010 年至 2019 年间，用户发表的评论数目及其相应给出的评价星级都呈快速上升趋势，这与互联网的发展息息相关；用户评论发表数量和平均评价星级都在 2018 年达到了顶峰，数量达到了 274939 条，平均评价星级高达 3.87 星。

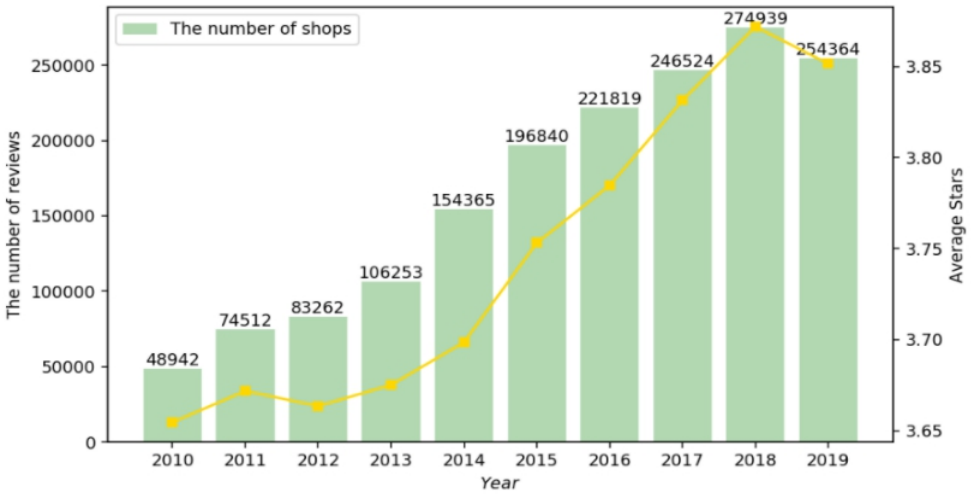


图 8 2010-2019 年拉斯维加斯餐饮商户用户评论总数及平均评价星级情况

除此之外，本文将评价星级为 1 星（最坏差评）和 5 星（最优好评）的具体评论文本经过词性标注，筛选出形容词和副词，得到相应的词云图，如图 9 (a)、(b)所示。可以观察得出，一星评论文本中重复出现了“n’t good”、“bad”、“never back”、“horrible”、“terrible”等贬义词，而五星评论文本中则重复出现出了“really good”、“great”、“definitely back”、“awesome”、“amazing”、“delicious”等褒义词。

(a)一星评论文本　　(b) 五星评论文本

图 9 一星和五行评论文本词云图（形容词&副词）

3 模型的建立与评价

3.1 随机森林确定影响因子权重

本文希望通过“预训练”的手段，初步判断特征对商家口碑的影响。具体，本文构建了随机森林模型，计算并得到各个特征在分裂时，对商户评价星级的影响权重。

3.1.1 特征工程

考虑影响商户评价口碑的因素，首先是商户的自身属性，如价格、营业时间、环境等等。同时，商户的历史评价口碑，对未来的消费评价也有着非常重要的导向性。除此之外，商户与商户之间也会相互影响，具体来说，那些相距较近的商户们，或是产生竞争关系、或是商户聚集形成商圈吸引更多客流量形成相互促进关系，这些都势必会对周围其他商户的消费评价产生影响。基于以上分析，随机森林模型所采取的变量选择如表 2 所示:

表 2 随机森林变量选择

因变量	数值型	Stars_log_weighted_2019（2019 年的 log 加权评论平均评价星级）
自变量	数值型	Review_count、Week_open、Hour_daily、Days_open、Neighbour_100、Neighbour_200、Neighbor_500、Neighbor_1000、Star_b2019_log_weighted
	布尔型	Wifi、Business_parking、Out_door、Credit_card、Group、Reserve
	等级型	Noise、Price

注：具体变量含义请见表 1

3.1.2 训练集、测试集划分以及参数设定

经过实验，设定树的数量为 100，树的最大深度为 15，每棵树用于分裂的特征个数由模型自动选择，训练集、测试集随机划分，比例为 7:3。

3.1.3 特征重要性权重

完成模型训练后，可得各个变量对预测的重要性程度，如表 3 所示。

表 3 随机森林模型各特征重要性权重

变量名称	Star_b2019_log_weighted	Review_count	Neighbor_1000	Hour_daily	Neighbor_500	Neighbor_200
权重系数	59.92%	8.65%	5.75%	5.10%	4.81%	4.62%
变量名称	Neighbor_100	Price	Business_parking	Group	Wifi	Noise
权重系数	4.31%	1.87%	1.06%	0.72%	0.68%	0.62%
变量名称	Days_open	Out_door	Reserve	Credit_card	Week_open	Eat_drink
权重系数	0.49%	0.46%	0.40%	0.30%	0.26%	0.00%

由表 3 可知，影响当前餐饮商户网络口碑最重要的因素是其历史的评价星级，影响权重高达 60.1%，其次为商户的评论量。同时，商户周围的商户数、每周开业时间也极大影响着消费者对商户的评价。再者则是商户自身属性如价格、营业时间、环境等因素。

3.2 NAR 模型

为了进一步定性探究各变量对当前餐厅商户网络口碑的影响作用，本文采用了 NETWORK VECTOR AUTOREGRESSION（The Annals of Statistics, Ann. Statist. 45(3), 1096-1123, (June 2017)）模型来兼顾商户历史口碑影响、邻近商户影响和商户自身属性影响这三类影响因素来源。

3.2.1 NAR 模型简介

在 NETWORK VECTOR AUTOREGRESSION 模型中，考虑N名用户形成的网络$A = [a_{ij}]$，每个用户有着独特的属性变量 Z，并各自生成一系列时间数据 Y_{it}。其中，NAR 模型认为 Y_{it} 由该机制生成：

$$Y_{it} = \beta_0 + Z_i^{\top}\gamma + \beta_1 n_i^{-1} \sum_{j=1}^{N} a_{ij} Y_{j(t-1)} + \beta_2 Y_{i(t-1)} + \varepsilon_{it}\,. \qquad (1)$$

式（1）中，β_0 为截距项，γ 为用户属性系数，β_1 为网络系数，β_2 为自回归系数。总而言之，NAR 模型中，用户的表现同时受到用户自身前期表现、与用户在网络中相连的其他用户前期表现、以及用户自身属性，这三类因素的影响，完美契合 Yelp 实际数据应用的目的。

3.2.2 模型构建

3.2.2.1 商户网络口碑时间序列数据

在 3.1 模型选取变量的基础上，考虑到数据的时效性以及连续月度数据的完整性，本文进一步选取了 2017 年 1 月至 2019 年 11 月间，共连续的 35 个月份、873 支商户的月度平均

评价星级数据。同时，为了兼顾星级与评价内容（文字内容较多的评价质量更高，影响力和参考性更强），在反映了商户的评价口碑的同时，更反映出该评价口碑的实际影响力，本文构建了每月相应的 log 加权星级作为 Y_{it}，即，假设当月共K条评论：

$$Y_{it} = \frac{\sum_{k=1}^{K} star_{itk} * \log(1+n_{itk})}{\sum_{k=1}^{K} \log(1+n_{itk})} \quad (2)$$

式（2）中，$star_{itk}$ 为商户i 在t月收到的第k条评论的星级，n_{itk}为该条评论的字数。

3.2.2.2 商户间关系网络

本文以商户间距离作为网络边权重的参考依据，构建商户间网络$A = [a_{ij}]$：

$$a_{ij} = \begin{cases} \frac{1}{d_{ij}+1} & d_{ij} < 2 \\ 0 & d_{ij} \geq 2 \end{cases} \quad (3)$$

式（3）中，d_{ij} 为商户 i 与商户 j 之间的距离（千米）。考虑到多数商户间距离过远，而消费者的用餐选择大多采取就近原则，因此我们对该网络进行了截断，对相距超过 2 千米的商户，认为商户间缺少有效关联性，网络取值 a_{ij} 截断为零。

3.2.2.3 商户自身属性

综合考虑随机森林模型结果与数据完整性，本文选取了 Hours_daily、Days_open、Wifi、Business_parking 等九个变量作为商户自身属性。考虑到不同档次商户的市场定位、营销策略以及目标消费群体都会有较大不同，为进一步探究其中的影响机制与差别，针对性的提供建议，本文以人均消费价格评级为标准划分了中低档商户（人均消费价格评级为 1 和 2，占比 84.88%）和高档商户（人均消费价格评级为 3 和 4，占比 15.12%）。并分别进行了模型拟合。

3.2.3 模型结果

经拟合可得到如表 4 的结果：

表 4　NAR 模型回归结果

	β_0	β_1	β_2	Price	Days_open	Week_open
中低档商户	2.446***	0.053***	0.462***	-0.068***	-0.069***	-0.057***
高档商户	2.219***	-0.100***	0.334***	0.089***	0.140***	-0.157***
	Hour_daily	Wifi	Business_parking	Group	Reserve	Credit_card
中低档商户	-0.018***	0.029***	0.195***	-0.033	-0.036***	-0.032
高档商户	-0.010***	-0.014*	\	-0.189***	0.130***	\

注：样本中高档商户均支持停车和信用卡，故略去；后缀*** 为 1%显著，*为 10%显著

表 4 结果显示：（1）所有商户都有非常显著的较大的正自回归系数，说明历史口碑对未来口碑具有重要的影响，且中低档商户相比于高档商户自回归系数更高，说明其口碑会更加影响未来消费评价。（2）中低档商户网络系数为正，而高档商户为负，说明中低档商户的评价星级之间存在着显著的互促关系，而高档网络间则是强烈的竞争关系。（3）关于商户自身属性：①对于价格范围，中低档商户系数为负显著，而高档商户为正显著，说明对于中低档商户来说更高的价格会降低商户，而对于高档商户来说，更高的价格会带来更好的评价。②对于营业天数及时间，可发现，营业时间过长会对网络口碑有显著的抑制作用。③中低档商户提供 Wifi 和停车位对网络口碑的提升作用都显著为正，但接受预定显著为负，且聚餐、信用卡作用都不显著 ④高档商户接受预定对提升口碑具有显著的促进作用。

4 商业应用与总结

从前面的模型结果中可以合理推断，中低档商户更容易聚集形成商圈，从而吸引更大的客流量，共同促进消费，并且中低档商户的消费评价十分依赖往期消费口碑，样本数据显示，每家中低档商户平均受到同一用户的 1.71 次评论。同时，中低档商户的消费群体多为价格敏感型，低价格往往可以换来更好的评价。与此同时，更多的营业时间并不能带来更好的消费评价，过多的无效营业时间只会带来更高的运营成本并且降低服务质量，使得评价变差。中低档商户可以通过改善环境设施，如提供免费 Wifi、停车位等方式来提升消费评价，但增大客位、提供信用卡、预定意义不大。而对于高档餐厅，商户网络间的竞争效应尤为强烈，同时更高的价格能够带来更好的评价，其消费评价并没有过多依赖往期消费口碑，样本数据显示，每家高档商户平均受到同一用户的 1.25 次评论。这说明高档餐厅的消费群体多为价格不敏感型，并且更加注重单次的消费体验，因此高档餐厅可以在采用更高的价格来保证服务品质，而停车位、信用卡消费等基础设施更是必备条件。同时高档餐厅更多是小规模用餐，而多人聚餐往往会带来更多的喧扰嘈杂，破坏高档餐厅的气氛。预定是消费者体验的先决保证，所以预定对高档餐厅来说必不可少。

总而言之，中低档餐厅更适合做好基础设施，通过聚集形成商圈的方式吸引更多客流量，打价格战以量取胜。而高档餐厅则首先需要保证服务质量，彰显自身的品味和档次，用更高的价格获取利润。