加盟餐饮品牌的选址预测

——以西式快餐品牌 T 店为例

势在必得队

摘要

西式快餐依托外卖平台逐步扩大市场规模，却也面临产品同质化的困境，某家中西结合的本土加盟式汉堡店 T 店得益于充分融合中式特色正在迅速扩张、占领西式快餐市场。本文通过 ArcGIS 对深圳市南山区、福田区进行栅格划分，基于研究区域现有西式快餐门店的经营数据与周边数据，用随机森林对 T 店加盟店铺选址概率进行预测；并通过线性模型、XGBoost、随机森林等多种方法建立回归模型，对初步选址进行销量预测，实现 T 店加盟店铺在深圳市南山区与福田区的精确选址方案。

一、背景介绍与研究问题

（一）背景介绍

1、行业分析

新冠疫情对餐饮堂食带来巨大冲击，而外卖的市场规模则在稳步扩大。截至 2022 年底，我国网上外卖用户规模已有 5.21 亿，外卖餐饮行业市场规模同比增长 16%。这一背景下，标准化程度高、适合外带外卖的西式快餐也迎来了发展契机，经历 2020 年受疫情影响而出现的负增长后，2021 年西式快餐市场规模达到 2800.7 亿、同比增长 13.5%，一改 2017-2019 年间不足 10%的疲态增速。据 QY Research 预测，2025 年西式快餐市场规模将达到 4996.5 亿元。可见，西式快餐市场规模仍在持续迅速增长，具有较大的市场空间。

图 1　2016-2025 年中国西式快餐市场规模现状及预测

与此同时，西式快餐市场赛道长期以来却面临着产品同质化严重、竞争加剧的困境。目前西式快餐的各大领头品牌均存在核心食材单一、产品工艺宣导空白、产品结构同质化严重、门店老旧等问题。困境与压力下各个品牌都在从不同方面尝试创新，尽管西式快餐领头品牌 K 店与 M 店仍占有相当大一部分的中国西式快餐市场份额，但这一数字已分别由高点时的 40%、17%降至 2021 年的 12.0%、6.7%，这是多家西式快餐品牌尝试打破同质化、分占市场的结果。如第一批崛起的本土西式快餐中，H 店通过价格战的方式迅速占据市场，在全国已

有 20118 家门店，与门店数量分别为 9205 家和 5634 家的 K 店与 M 店形成差异化竞争。但随着越来越多品牌的加入，主打低价消费的下沉市场的竞争也愈演愈烈，仅靠低价策略来复制 H 店的成功已成天方夜谭，想要在西式快餐这个赛道中分一杯羹需要更多新思路。

2、T 店品牌分析

主打中国特色的 T 店作为西式快餐的后起之秀，凭借独特的风味迅速占据了西式快餐市场的一席之地。T 店在 2020 年爆红，仅 2022 年开店数量就超过 1500 家，一跃成为西式快餐中门店数量仅次于 H 店、K 店、M 店的领头品牌。不同于传统的西式汉堡，中西结合的 T 店的汉堡口味融合中式经典餐饮特色，产品创新主推手感汉堡胚，店铺装修以国潮图案为主，一定程度上打破了西式快餐同质化的困境。同时，T 店以 20.36 元的人均消费打开了下沉市场，据极海数据显示，约 71.29%的门店开在二三四线城市，仅有 22.78%的门店开在一线和新一线城市，因此 T 店在一线城市仍有较大的发展空间。

图 2　T 店在全国门店现状

（二）研究问题

在外卖行业蓬勃发展、西式快餐市场规模持续增长的当下，我们选择外卖订单量长期在全国名列前茅的深圳市及正处于快速扩张期的西式快餐品牌 T 店进行数据分析，聚焦写字楼和高校密集、年轻人众多、生活节奏快速的深圳市南山区福田区，思考如何依托外卖平台、通过基于历史数据从而科学地对 T 店进行选址来充分打开一线城市市场，实现利润最大化。

二、数据的说明与描述

为了进一步进行分析，本文收集了大量相关数据，以保证能够在基于充足的历史数据和适合的特征选择的情况下进行科学地选址决策。本文所用数据的来源及简要说明如表 1 所示。首先将深圳市南山区与福田区进行 1km×1km 网格级别划分，便于为研究区域空间位置的比较提供统一尺度及整合数据。而后通过下述方法获得相应数据并进行数据清洗和数据处理，将数据按照栅格进行整合并赋给每个栅格，并用热力图对数据进行展示，处理后的数据

为各个栅格的统计特征，该数据将用于对 T 店门店初步选址进行概率预测。最后选取的特征为：人口密度、办公楼及高校密度、路网密度、夜间灯光指数、房价、外卖月销量，部分特征的计算公式如下：

$$\text{路网密度} = \frac{\text{道路长度}(km)}{\text{道路所占区域面积}(km^2)}$$

$$\text{夜间灯光指数} = \text{图像灰度值}^{3/2} \times 10^{-10}$$

$$\text{外卖月销量} = \sum_i \frac{\text{店铺}i\text{周边}2km\text{范围与栅格相交面积}(km^2)}{\text{栅格面积}(km^2)} \times \text{店铺}i\text{外卖月销量}$$

$$\text{办公楼高校密度} = \frac{1}{\text{搜索半径}(km)^2} \sum_i \frac{3}{\pi} \times \left(1 - \left(\frac{\text{办公楼或高校}i\text{与栅格中心间距}(km)}{\text{搜索半径}(km)}\right)^2\right)^2$$

表 1　数据来源与说明

数据类型	数据来源	数据说明
POI 数据	高德地图 API 数据开放接口(https://lbs.amap.com/)	通过 Python 网络爬虫技术调用高德地图 API 爬取
人口数据	WorldPop 全球高分辨率人口计划项目数据集 (www.worldpop.org)	门店布局与人口空间分布存在明显的相互吸引效应，人口规模越大，消费需求的积累越容易产生，市场潜力越大
路网数据	OpenStreetMap(https://www.openstreetmap.org)	路网密度是衡量区域交通可达性的重要指标之一，商业集聚程度高的区域一般路网密度高
夜间灯光数据	珞珈一号 (http://59.175.109.173:8888/app/login.html)	夜间灯光与区域生产总值 GRP 存在较高的相关性，门店布局往往选择经济发展程度高的区域
房价数据	链家(https://www.https://sz.lianjia.com/)	通过 Python 网络爬虫技术爬取。城市房价与居民平均收入水平正相关，居民收入水平越高，购买力越强，房价就越高
店铺外卖数据	美团外卖(https://h5.waimai.meituan.com/)	通过 Python 网络爬虫技术爬取。反映地区外卖行业发达程度和周边居民生活习惯
行政区划数据	深圳市政府数据开放平台 (https://opendata.sz.gov.cn/)	表征深圳市福田区与南山区行政区划边界，通过矢量底图掩膜提取出各因子栅格

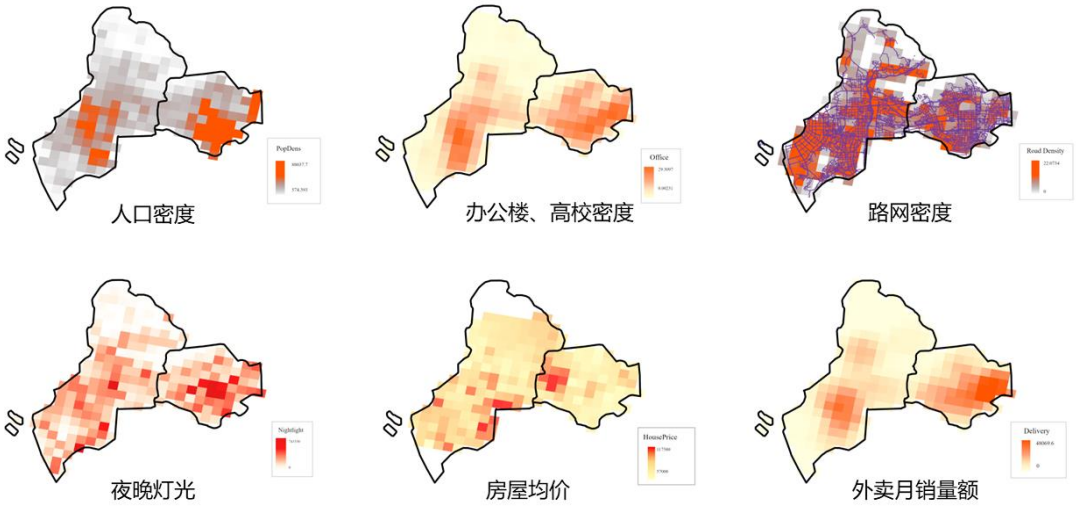

图 3　ArcGIS 统计特征可视化

由于 T 店的主要特色产品为汉堡，因此应着重参考主要产品同为汉堡的西式快餐店（下文简称“汉堡店”）的历史数据。得到初步选址的概率预测后，进一步进行基于历史数据的销量预测，从而实现精细化选址。在进行销量预测时，应以汉堡店为单位构造特征，因此假设 T 店初步选址位置为所得栅格的中心位置。以下为基于原始数据对每家店铺构建的数据指标及历史数据训练集对应指标的简单描述性统计。

表 2　数据指标说明及描述性统计

		变量名	平均数	标准差	最大值	最小值	中位数
因变量		汉堡店月销量	1682	2141.04	9999	49	900
解释变量	经济特征	汉堡店人均消费额	29.98	10.52	86.00	10.00	30.00
		房屋均价	81434	6280.513	102256	69670	80165
		夜间灯光指数	65677	34177.17	161779	7384	55356
	区域特征	人口密度	26899	12757.16	56395	1535	25160
		办公楼、高校密度	14.88	5.18	25.58	0.95	15.41
		路网密度	7.29	1.62	11.62	2.28	6.96
	竞争情况	2km 内竞争对手数量	16.65	9.74	37	0	16
		2km 内汉堡店密度	0.0536	0.0299	0.1413	0.0031	0.0522
		2km 内竞争对手平均距离	1.17	0.24	2.00	0.61	1.18
		2km 内竞争对手销量	24595	13168.36	49047	0	25099
		2km 内汉堡店销量比	0.0977	0.0432	0.4348	0.0402	0.0861

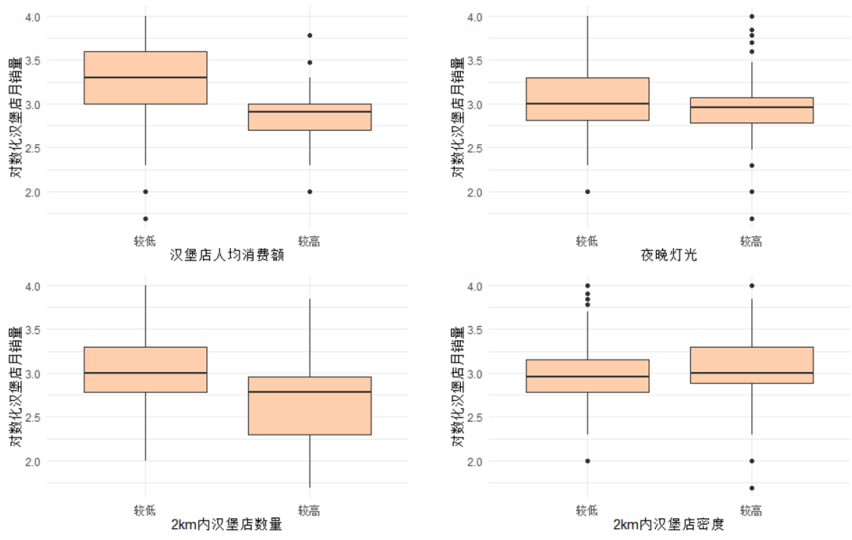

图 4　部分解释变量与因变量箱线图

通过均值、标准差、最值、中位数等描述性统计量对各个变量的分布进行简单考察，发现大部分变量均值与中位数相差较小，且最值都分布在均值的三倍标准差范围内，可以初步说明大部分变量分布偏度不高、异常值较少。接下来将部分变量根据中位数划分为两组，并通过箱线图来观察各解释变量与因变量之间关系，初步判断个解释变量对因变量是否存在影

响及影响方向。从图 4 中可以看出人均消费、和 2km 内竞争对手数量均与对数化月销量呈现负向关系；2km 内汉堡店密度则与对数化月销量呈现正向关系；而夜晚灯光的两个组别间虽然呈现出一定的差异，但该数据异常值点较多，因此难以判断与因变量的影响方向。

三、模型的建立与评价

下面我们分三个环节进行 T 店选址。第一步是对缺失数据的处理，由于部分数据的可获得性不强，例如房价数据，因此本文首先使用 R 中的 mice 包对缺失数据进行填补，使用的方法是多重插补法；第二步是根据数据的可获得性和质量水平，结合现有研究成果，以栅格为单位选取了 6 个特征，然后以栅格内汉堡店的有无作为因变量构建随机森林二分类器，先筛选可能开汉堡店的网格；第三步是基于随机森林分类的结果，提取经济、区域、竞争三个类别的 11 个特征，然后以对数化汉堡月销量为因变量建立回归预测模型，筛选出销量最高的 3 个选址。

（一）随机森林分类器模型

本文以 228 个栅格的数据作为样本集，一共具有 6 个特征向量，以是否具有月销量 1000 以上的汉堡店作为因变量，尝试使用 XGboost 和随机森林两种方法来构建二分类模型，然后用混淆矩阵和 ROC 曲线的下面积值（AUC 值）来评估模型的分类性能。两个模型的 AUC 值分别是 0.7118 和 0.8854，所以我们选择结果更优的随机森林分类器。

1、实验过程

基于随机森林分类器对深圳市南山区福田区的 T 店的选址进行概率预测的实验过程为：

（1）选取影响 T 店的 6 个特征并计算每个网格中相应的特征值用以构建特征矩阵；

（2）提取所有样本以构建训练数据集，由于正样本数量过少，因此我们通过计算正负样本的比例构造样本权重，然后作为随机森林模型 TuneRF 函数和 TrainForest 的 cweights 参数，这样可以平衡正负样本数量，从而使得模型训练更加稳健。

（3）从训练数据集中随机选取 70%的数据用于训练，剩余 30%的数据用于验证，然后通过数据划分、交叉验证、参数调优、平衡模型训练复杂度等方法选取最佳模型，并输出模糊矩阵和 ROC 曲线并计算 AUC 值进行模型评估；

（4）将特征矩阵导入预先训练好的模型进行分类预测，得到每个网格布局 T 门店的概率并进行适宜性制图。

2、模型可靠性评估

传统衡量机器学习算法性能的指标通常选择测试集的 Kappa 系数和精度，但该评价方法往往会忽略机器学习模型后验概率大小程度，故不能反映机器学习算法的真实性能。而

ROC 曲线的下面积值(AUC 值)不仅能有效检验模型的分类性能，还能度量机器学习算法的后验概率和排序性能，在机器学习算法可靠性评估中应用广泛。基于以上分析，本文综合使用混淆矩阵、ROC 曲线及 AUC 面积来评估模型性能。

（1）混淆矩阵：基于混淆矩阵选取精确度、召回率和F_1分数值评估模型性能，其中F_1分数值为精确度和召回率的谐波平均值，其值在[0,1]之间，越接近 1 表示分类器性能越好。

（2）AUC 面积：由于 ROC 曲线无法直接作为分类器的评价指标，故采用 ROC 曲线对应的下面积（AUC 值）对模型性能进行评估。

本文使用 R 自带的 TunRF 函数对随机森林分类器的参数进行调整，然后计算测试集的F_1分数值以及 AUC 值。实验的结果为F_1分数值为 0.8913，且如图 5 所示，ROC 曲线呈现明显的左上突趋势，离纯随机分类器的 ROC 曲线（图中对角线）较远，AUC 面积为 0.8854，以上分析结果表明，经过调参之后的模型整体性能表现优秀。

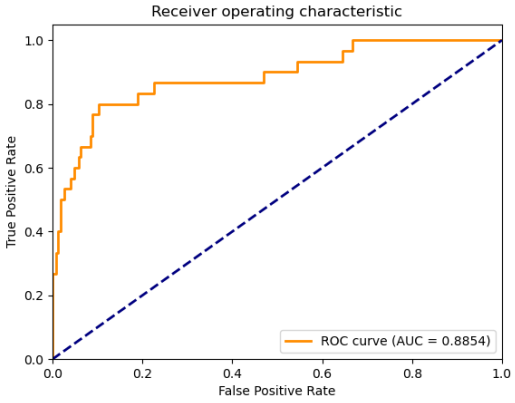

图 5　ROC 曲线及 AOC 值

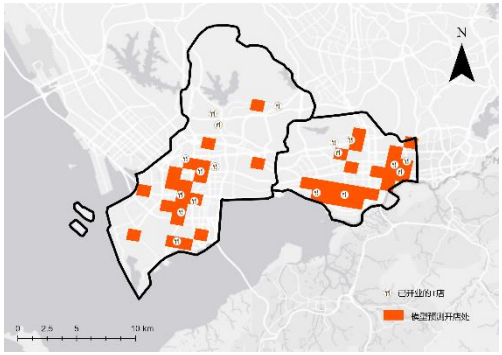

图 6　T 门店的预测选址

下面我们将原始数据导入预先训练好的模型中并进行预测，得到汉堡门店在深圳市福田区和南山区的预测选址，然后用 ArcGIS 在地图上画出预测选址如图 6 中橙色栅格所示。从预测选址的网格可以看出来，汉堡店的预测选址比较集中且大多数都覆盖已有的地址，下面我们将利用 T 店的特征预测在这些选址中开 T 门店的月销量。

（二）基于预测栅格的回归模型精确选址

为了得到精确的 T 门店选址，并深入挖掘影响 T 外卖销量的显著因素，本报告选取所有汉堡店为样本，然后以 11 个特征为自变量，销量为因变量。分别用随机森林和线性回归建立了模型，最后选择了在测试集上均方误差最小的线性回归。

线性回归的实验过程如下：

（1）首先对所有变量建立线性回归全模型，可以得到调整的 R 方为 0.7210，P 值为

2.2e-，这说明全模型的似然比检验高度显著，因此在考虑的所有因素中，至少有一个是对汉堡店外卖销量有显著影响的。

（2）下面我们根据 AIC 准则对变量进行筛选，最后的结果如下表所示：

表 3　AIC 模型回归结果

变量名	回归系数	标准差	P 值	备注
截距项	7.410	0.170	0.000	
汉堡店人均消费额	-0.023	0.004	0.000	
夜间灯光指数	3.601e-06	0.000	0.049	
2km 内竞争对手数量	-0.297	0.002	0.000	
2km 内汉堡店密度	86.900	5.510	0.000	
2km 内竞争对手销量	9.013e-06	0.000	0.126	
全模型检验	P 值<0.001			调整 R 方=0.728

汉堡店人均消费额和 2km 内竞争对手数量的回归系数显著为负，这个实验结果是符合实际情况的，人均消费额与销量成反比，同时竞争对手数量上升也会使得销量下降；夜间灯光指数的回归系数为正，说明它与销量成正比，而夜间灯光指数为经济特征，说明一个地区的经济水平会影响汉堡店的销量；2km 内汉堡店密度的回归系数也为正，说明汉堡密度与销量也成正比，这符合产业集聚的原理，产业集聚可以提高产业的竞争力，因此汉堡店的集聚也能在一定程度上提高汉堡店的月销量。同时我们注意到 AIC 准则筛选变量时，筛选掉了所有的区域特征，这可能是因为我们选取的福田区和南山区都是深圳市的主要商业区，区域发展比较一致，所以对销量不具有显著影响。

（3）最后我们将随机森林分类器筛选出来的栅格进行精确选址，将特征变量输入 AIC 筛选后的模型，然后将预测得到的销量输入 ArcGIS 进行画图，在考虑 T 店的 1km 内不能开店的连锁规定下得到的精确选址如图 7 所示。

图 7　T 门店的精确预测选址

五、商业应用与总结

基于初步选址预测分类器与销量预测模型，我们设计了一款对加盟店铺选址工具，如图 8 所示。该系统可以给出目标城市目标加盟店铺的初步选址布局，并给出初步选址位置的销量预测，便于加盟商进行最优的店铺选址。

图 8　商铺选址工具

本报告基于西式快餐依托外卖平台而逐步扩大市场规模的现状，立足于 T 汉堡门店的选址，考虑为加盟店铺选址的问题。首先为了给两个区空间位置的比较提供统一尺度，对空间进行栅格处理；接着以栅格为单位选取 6 个特征，用随机森林分类器进行初步选址；然后综合考虑区域特征、经济特征、竞争特征三个方面的 11 个特征建立了线性回归模型，并基于 AIC 准则筛选出五个最为显著的变量，对分类器选址地店铺的月销量进行预测，最后综合考虑 T 店的连锁保护规定和销量最高的目标，选出了三个推荐加盟选址。

本报告依然存在诸多可改进之处，主要包括：（1）由于数据获取途径的限制，选取的特征变量集以及样本数量有待扩充；（2）所选研究区域同质化较高，可以扩充研究区域从而进行更加准确的预测。如果对这两点进行改进，本报告在预测的准确性和解释性上应能取得较大进步。