

应用机器学习对外卖含糖饮料消费的研究^{*}

翟倩倩 (浙江大学中国农村发展研究院 杭州 310058)

蒋 玉 (西南财经大学中国西部经济研究院 成都 611130)

张家铭 (普渡大学农业经济系 西拉法叶 47907)

胡安·埃斯特班·格瓦拉 (普渡大学统计系 西拉法叶 47907)

王 红 (普渡大学农业经济系 西拉法叶 47907)

摘 要: 含糖饮料因其高热量而导致慢性疾病,引起的健康问题得到全球的关注。在外卖平台迅速发展的大背景下,本文基于中国六个城市的外卖平台的销售数据,运用机器学习方法识别外卖含糖饮料消费动态,并对价格折扣、糖税等模拟情景进行了销售变动分析。研究发现,第一,有监督机器学习模型中的神经网络相较于传统计量回归模型更加适用于消费趋势预测分析。第二,模型预测结果表明,当价格降低 10% 时,总销售额将比原始销售增加 0.49%,说明含糖饮料消费富有价格弹性;此外,当价格上升 10% 时,总销量额将比原始销售下降 0.17%,表明通过税收手段来减少含糖饮料的消费是可行策略。第三,外卖含糖饮料的销售情况存在时空差异,长三角地区外卖含糖饮料的销售量高于其他地区,周中的销售量高于周末时段,下午的销售量是一天的高峰。本文研究对于建立降低含糖饮料的过度消费的路径与机制,引导消费者健康饮食选择,推动外卖服务行业的可持续发展具有重要意义。

关键词: 含糖饮料; 外卖; 机器学习; 神经网络

一、引言

食品和饮料的消费与健康息息相关,不健康的饮食导致一系列健康风险,其中,超重和肥胖被认为是最需要关注的也是全球增长最快的风险问题(Afshin 等,2017),被称为全球性流行病(WHO, 1998)。预测数据表明,到 2025 年,全球约有 2.06 亿 5~19 岁的儿童和青少年将面临肥胖问题(Lobstein 等,2019),中国的肥胖儿童数量将居于全球首位(Afshin 等,2017)。越来越多的证据表明,含糖饮料的过度消费对肥胖率的增长有显著贡献

(Hu 等,2010; 全世文,2021)。

含糖饮料包括碳酸饮料、果汁饮料、运动饮料、能量饮料以及新近流行的奶茶类饮料,大多含有蔗糖、高果糖、玉米糖浆等高卡路里甜味剂。世界卫生组织(WHO) 建议每天摄入的添加糖不应超过总能量的 10%(WHO,2015),含糖饮料在饮食中占据了添加糖的主要来源。与高收入和低收入国家相比,含糖饮料的消费在中高收入和中低收入国家中更高(Singh 等,2015),而含糖饮料销售量最显著

^{*} 项目来源:教育部人文社科重点研究基地重大项目“多重冲击与韧性食物系统研究”(编号:22JJD790079),四川省社会科学基金青年项目“数字经济背景下四川省粮食品牌发展路径与政策保障研究”(编号:SCJJ23ND418),教育部人文社会科学研究一般项目“食品公众政策与健康食品消费:基于含糖饮料消费的实证研究”(编号:21YJA790013)。蒋玉为本文通讯作者

的增长出现在中国、智利、泰国和巴西(Popkin 等, 2016)。近年来饮食环境的变化,特别是后疫情时代在线餐饮配送(OFD) 平台的迅速扩张(Kumar 等, 2021), 更是助力了含糖饮料消费的增长。

公众逐渐认识到罐(瓶) 装的碳酸饮料属于不健康食品, 新鲜制作的奶茶和咖啡悄然赢得消费者喜爱。外卖配送行业的迅速发展, 使得购买行为不再受距离限制, 实现足不出户却能享用鲜制含糖饮料的便捷(de Freitas 等, 2020)。中国的 OFD 交易已经达到 5.44 亿, 市场规模达到了 8117 亿人民币, 在餐饮行业中的渗透率达到 17%(NBS, 2022)。遍布各地的外卖网络降低了消费者的含糖饮料购买交易成本, 消费者在手机或网页上获取外卖信息如美团和饿了么等, 选择偏好的饮料后在线支付, 利用配送服务收货。外卖服务整体环节便捷、灵活、时效高。截至 2018 年, 中国奶茶市场的零售销售额超过了 500 亿人民币, 其中鲜制奶茶销售额超过了 400 亿人民币。从 2016—2018 年, 中国奶茶外卖订单的季度复合增长率达到了 38%。

作为典型的新兴经济体, 中国目前正在经历着一场双重转型, 表现为外卖平台的迅速扩张和饮食模式的显著转变——从传统的低添加糖的中式饮食向更西式的饮食风格演变(Dai 等, 2022; Gao 等,

2020; Zhang 等, 2023)。这一转变导致了饮食营养结构的不平衡, 增加了中国人慢性疾病的患病率(樊胜根等, 2023)。考虑到这种情况, 迫切需要对外卖中的含糖饮料消费进行研究, 识别其消费趋势。这对于如何制定干预措施, 引导消费者提高饮食质量和改善整体营养健康状况有关键指导意义。

然而已有研究主要集中于分析 OFD 服务的感知、态度和使用意愿(Yeo 等, 2017; Roh 等, 2019; Cho 等, 2019), 以及消费者对外卖服务的满意度和忠诚度(Suhartanto 等, 2019) 等, 尽管有部分学者开始关注外卖服务对饮食质量、食品安全和营养摄入的影响(Maimaiti 等, 2018; Gupta, 2019; Stephens 等, 2020), 极少有研究探索外卖平台中的含糖饮料消费及其潜在影响。

因此, 本文基于中国一家外卖平台获取的高频数据, 应用机器学习技术, 旨在探索外卖服务背景下含糖饮料消费趋势与特征, 并尝试建立降低含糖饮料过度消费的可行路径。具体而言, 本文采用前沿的机器学习方法, 利用历史销售数据构建模型, 识别消费模式, 并随后利用这些信息模拟价格折扣、征收糖税等情景下的含糖饮料的消费趋势, 以系统分析外卖服务背景下含糖饮料消费的动态, 为相关政策制定提供建议。

二、机器学习应用现状及前景分析

(一) 机器学习的发展历程

从理论层面来看, 机器学习(Machine Learning) 是人工智能(Artificial Intelligence) 的一个重要分支(Bell, 2020; Taddy, 2018)。它在医疗保健、金融服务、自动驾驶和市场营销等多个领域展现显著优势(Jordan 等, 2015)。在过去的几十年中, 机器学习不仅在实际应用上取得了快速发展, 其学术进展也相当显著。许多经济学家认为, 机器学习将对经济学的未来发展产生广泛且深远的影响(Athey, 2018; Varian, 2014; Belloni 等, 2014; Brynjolfsson 等, 2014; 陶旭辉等, 2023)。此外, 机器学习的兴起也促进了计算社会科学的发展, 形成了一种全新的研究范式和交叉学科。在这个新兴趋势下, 数据、算法和算力成为了基本的要素(Chen, 2018; Mason 等, 2014)。

目前, 机器学习的研究主要应用了三种核心方

法, 分别是有监督机器学习(Supervised Machine Learning)、无监督机器学习(Unsupervised Machine Learning) 和强化机器学习(Reinforcement Machine Learning)。其中, 有监督机器学习是应用最为广泛的方法(Hastie 等, 2009)。这种方法通过使用被标注的样本数据来训练模型, 其核心在于建立特征值(即经济学中常用的解释变量) 与标签值(即被解释变量) 之间的对应关系, 并构建出最优模型(Kotsiantis et al, 2017), 从而对目标标签值进行有效的估计和预测。与传统的计量经济学中的回归方法相比, 有监督机器学习方法允许模型自动从数据中学习, 不需要设定数据的分布和假设, 也不需要设定解释与被解释变量间的函数关系形式, 因此可以更加灵活、有效地选择函数形式, 从而避免过度拟合。此外, 有监督机器学习更注重预测结果, 而非参数估计。

Hastie 等(2009)学者提出了多种有监督机器学习的映射形式,包括决策树、决策森林、逻辑回归、支持向量机、神经网络、核机器和贝叶斯分类器等。在这些方法中,与深层网络相关的领域影响尤为广泛。这些网络由多层结构组成,每层都包含一个基于简单计数化函数计算的阈值单位(Bengio, 2009; Schmidhuber, 2015)。神经网络是目前最广泛使用、有效的监督机器学习方法,特别适用于处理时间序列、面板或空间数据等非线性关系(Storm 等 2020; 王川等 2008)。

无监督机器学习主要应用在非标注的样本数据中。它可以从视频、图像等没有标签值的数据中发现隐藏的模式、规律或关系,从而对数据进行聚类、降维等处理。无监督机器学习可以用来对数据进行预处理,在传统的计量经济学的回归分析之前检验数据之间的相关性(Hodeghatta 等 2017)。强化机器学习区别于有监督和无监督学习,它不仅依赖于已知的、固定的数据进行学习,而是在不断变化的环境中,通过与环境的互动和大量的试错学习,寻找能够产生最佳结果的有效路径(Taddy, 2018)。这使得强化机器学习特别适用于需要复杂、动态环境下的决策研究,如游戏、机器人控制和自动驾驶等领域。

(二) 机器学习在农业经济学的应用现状

机器学习技术已开始农业经济研究中发挥作用(于晓华等 2019)。Graskemper 等(2021)通过采用了无监督的机器学习方法,一项涵盖个人、农场和背景等广泛客观变量的大型调查,对德国农民结构的分析,从数据集中提取了有价值的见解。在另一项关注中国鸡蛋价格的研究中,机器学习技

术被应用于对区域零售鸡蛋价格进行聚类分析。研究人员通过将机器学习的结果与计量模型相结合,深入探讨了集群隶属的决定因素,为了解不同地区鸡蛋价格的影响因素提供了更加深刻的分析(Liu 等 2023)。在对消费者偏好的研究中,Ruiz 等(2020)采用机器学习分析了消费者行为。他们从 5000 多种产品的选择中提取信息并构建模型,揭示了各个产品之间的替代或互补关系。

机器学习方法在预测方面也显示出了强大的应用潜力。Zhao(2021)采用机器学习方法分析影响农产品期货价格的基本因素以预测农产品未来的价格走势。Gopinath 等(2021)利用七种主要农产品的历史交易数据,采用有监督机器学习和神经网络方法分别基于 2010 年和 2014 年的数据进行机器学习模型训练。他们的研究结果显示,机器学习模型在近期和长期贸易模式预测方面具有较高的相关性。

在农业经济学领域,也有研究使用无监督机器学习方法。Zeng 等(2023)基于中国经济、人口和社会融合的多维度指标,利用无监督机器学习方法识别城乡一体化的类型,确定了四个不同的特征聚类,即高水平城乡一体化、处于转变中的城乡一体化、低水平城乡一体化和处于滞后阶段的早期城乡一体化。然而在农业和食品经济消费领域,尤其是在餐饮外卖服务的背景下,关于机器学习的应用十分有限。其中,Shankar(2022)的研究强调了机器学习对提升外卖行业中消费者体验的实质性影响,说明将机器学习方法纳入对外卖服务的研究,可以成为推动外卖行业发展的一个重要因素。

三、外卖含糖饮料消费的应用案例分析

(一) 数据与变量

本文分析了中国一家用户量高的外卖平台的历史销售数据,时间跨度为 2023 年 1—7 月。该平台涵盖了几乎所有主要城市的外卖服务餐厅和餐饮服务商。研究选择了 6 个主要城市包括 3 个 1 线城市(北京、上海和广州)和 3 个 2 线城市(杭州、成都和天津)。本文仅使用所有交易数据的 1% 随机样本。

由于本文研究重点是含糖饮料,因此只保留了

含糖饮料的数据。在中国,除了餐食之外,由专门饮料店准备的饮料也是外卖订购的热门选择,星巴克便是其中一个热门店铺。近年来,珍珠奶茶或特调奶茶等饮品受到消费者的喜爱,许多专营饮料的独立奶茶店也提供外卖服务。本文在数据集中仅保留了饮料订单,总订单数量为 2339。

本文将含糖饮料分为三类,即奶茶、咖啡和其他饮料。值得注意的是,奶茶有各种各样的配料,包括水果、淀粉制品、乳制品、豆制品,尤其是糖,这

种类型的热量含量最高。咖啡包括普通咖啡和以浓缩咖啡为基础的特调咖啡,它们大多数也含糖。其他饮料指的是其他添加糖分的果汁类、奶昔类、茶类等鲜制含糖饮料。

每一笔订单是一个观察值,由一系列标识编号(ID)来索引,包括订单ID(ID-order)、店铺ID(ID-store)、城市ID(ID-city)、区域ID(ID-district)用于标识每个城市的某个特定区域)和用户ID(ID-user)。商店和配送目的地的位置通过纬度和经度进行索引(latitude-store、longitude-store、latitude-destination

和longitude-destination)。利用这些地理坐标作为位置变量,本文根据欧氏范数计算了每次配送的距离。不同配送距离的订单数量分布如图1所示,几乎所有订单都来自5公里以内,众数是来自附近1/4公里到1/2公里的店铺,而最密集的订单是1/4公里至1公里的范围,其余呈现出随距离增加而减少的趋势。值得一提的是,距离小于1/4公里的订单量未排在前列,这可能是由于店铺太近了,到店自取很方便,外卖配送显得不必要。

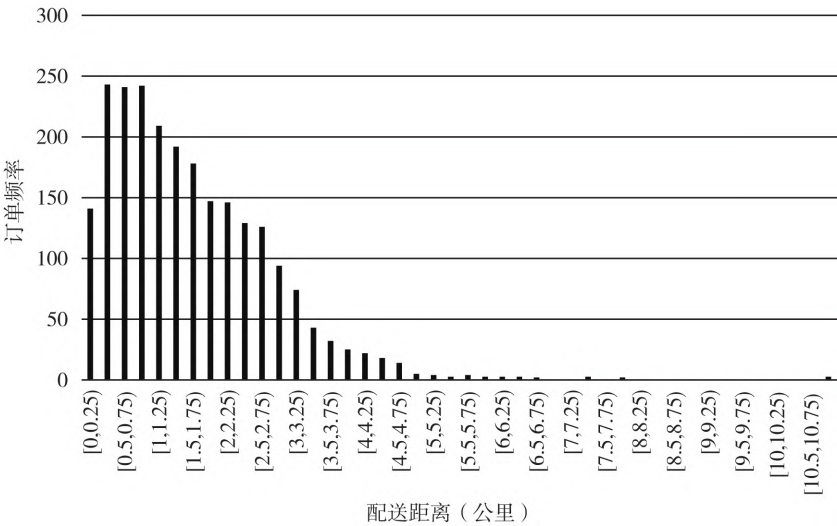


图1 配送距离的分布

每杯含糖饮料的价格主要集中在10~15元(见图2)。每笔订单的销售数量平均为3杯(见图3),标准差相对较低,表明销售数量分布相对较稳定。最大的一笔订单有54杯。本文创建了虚拟变

量来表示订购的饮料类型,即它们是否是咖啡、奶茶以及其他类型的饮料。如图4所示,大多数订单是奶茶,占63.32%;咖啡订单占最小份额的13.85%,其余为其他类型的饮料。

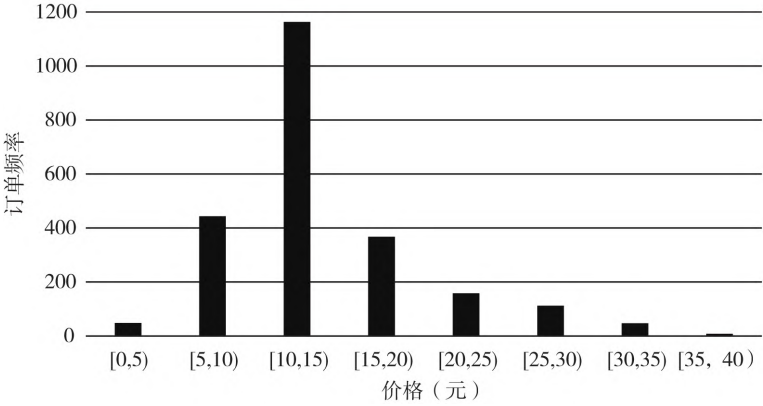


图2 价格的分布

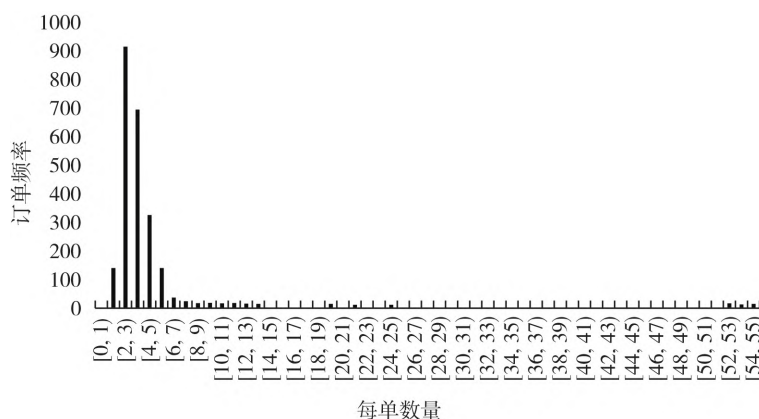


图 3 数量的分布

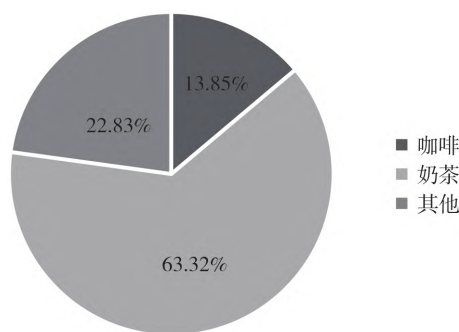


图 4 含糖饮料类型的比例

本文从日期信息中提取了时间特征,包括每笔交易的月份和星期。这些变量由月度虚拟变量和星期的虚拟变量表示,并有助于反映一年中的季节效应和一周的休息日效应。此外,本文将“时间”变量拆分为时段,分别为上午(6:00—11:00)、中午(11:00—14:00)、下午(14:00—15:00)、晚上(15:00—20:00)和夜间(20:00—次日6:00)。这种分类能够帮助对消费者的消费行为和时间偏好进行细致入微的分析。

本文还包括了表示六个城市中的 22 个区域的 22 个虚拟变量。区域虚拟变量以城区的名字命名。如表 1 所示,在分析中使用的每个变量的描述统计数据。结果显示,距离变量的均值为 1.58 公里,标准差为 1.13,表明平均而言,人们从离他们接收地点较近的商店订购含糖饮料。销售额呈现出均值为 28.05 元,标准差较高,表明数据集中销售额存在相当大的波动。最大销售额达到 404.10 元,平均价格为 13.79 元/杯。其相对较高的标准差意味着

在折扣或优惠之前产品定价存在较大差异,这也是饮料品种很多的结果。有大量免费配送的交易,配送费的最大费用为 25 元,较高的配送费主要是配送距离较远导致的。

(二) 机器学习模型的选择:神经网络(Neural Network)

鉴于本文所研究问题的特点,使用有监督的机器学习方法。又基于数据集的规模和复杂性,本文选择了深度学习模型(Deep Learning)。选择这种模型的理由在于深度学习能够处理大量数据并发现复杂的模式(Alzubaidi 等,2021)。本文使用了一个神经网络(Neural Network)结构,其中包括多个密集层,受到人脑感知处理的启发,这对于外卖平台系统中特有的表格和序列数据非常有效。这些网络仍然可以从深度学习提供的自动特征学习中受益,并且在捕捉和利用数据丰富多样的信息方面,与简单的机器学习模型相比具有显著优势。

表 1 描述统计表

项目	均值	标准差	最小值	最大值
销售额(元)	28.0501	23.3095	0.0100	404.1000
数量(杯)	3.2715	3.9653	1.0000	54.0000
配送费(元)	0.9300	2.0104	0.0000	25.0000
价格(元/杯)	13.7913	5.8077	0.2925	39.5000
折扣(元)	3.9470	2.7295	0.0000	19.2500
咖啡	0.1385	0.3455	0.0000	1.0000
奶茶	0.6332	0.4820	0.0000	1.0000
其他	0.2283	0.4198	0.0000	1.0000
距离(公里)	1.5842	1.1251	0.0002	10.8056
星期一	0.1270	0.3330	0.0000	1.0000
星期二	0.2044	0.4033	0.0000	1.0000
星期三	0.3087	0.4620	0.0000	1.0000
星期四	0.1308	0.3373	0.0000	1.0000
星期五	0.0770	0.2666	0.0000	1.0000
星期六	0.0556	0.2292	0.0000	1.0000
星期天	0.0966	0.2955	0.0000	1.0000
早晨	0.2069	0.4052	0.0000	1.0000
中午	0.1791	0.3835	0.0000	1.0000
下午	0.2984	0.4577	0.0000	1.0000
傍晚	0.2279	0.4196	0.0000	1.0000
晚上	0.0876	0.2828	0.0000	1.0000
上海市嘉定区	0.0265	0.1607	0.0000	1.0000
上海市普陀区	0.0321	0.1762	0.0000	1.0000
上海市杨浦区	0.0415	0.1994	0.0000	1.0000
上海市浦东区	0.0825	0.2752	0.0000	1.0000
上海市虹口区	0.0316	0.1751	0.0000	1.0000
上海市长宁区	0.0428	0.2023	0.0000	1.0000
上海市静安区	0.0162	0.1264	0.0000	1.0000
上海市黄浦区	0.0423	0.2014	0.0000	1.0000
北京市朝阳区	0.0389	0.1934	0.0000	1.0000
北京市海淀区	0.0073	0.0850	0.0000	1.0000
天津市南开区	0.0308	0.1728	0.0000	1.0000
天津市滨海新区	0.0385	0.1924	0.0000	1.0000
天津市西青区	0.0342	0.1818	0.0000	1.0000
广州市海珠区	0.0038	0.0619	0.0000	1.0000
广州市番禺区	0.0770	0.2666	0.0000	1.0000
广州市白云区	0.0175	0.1313	0.0000	1.0000
成都市武侯区	0.0410	0.1984	0.0000	1.0000
成都市青羊区	0.0410	0.1984	0.0000	1.0000
杭州市上城区	0.0539	0.2258	0.0000	1.0000
杭州市拱墅区	0.1958	0.3969	0.0000	1.0000
杭州市萧山区	0.0338	0.1807	0.0000	1.0000
杭州市西湖区	0.0710	0.2568	0.0000	1.0000
一月	0.0556	0.2292	0.0000	1.0000
二月	0.0851	0.2791	0.0000	1.0000
三月	0.0428	0.2023	0.0000	1.0000
四月	0.0650	0.2466	0.0000	1.0000
五月	0.0799	0.2713	0.0000	1.0000
六月	0.0534	0.2250	0.0000	1.0000
七月	0.6182	0.4859	0.0000	1.0000

注: 表中未标注单位的均为比率

在大多数情况下,研究中可观察到一个解释变量向量 $x_i \in R^d$ 和一个响应变量 $y_i, i=1, 2, \dots, n$, 而本文希望找到一个函数 $f(x)$, 以解释 y_i , 使得 $f(x_i)$ 的值尽可能接近 y_i 。线性回归是其中的一个特例, 其中假定 $f(x)$ 具有线性表达式 $f(x) = \beta^T x$ 。线性表达式显然是一种对一般函数形式极为限制的形式。因此, 非线性回归和非参数回归被提出作为捕捉数据中更复杂关系的替代方法。

机器学习中的神经网络提供了另一种拟合这样一个未知函数 f 的方法。单层神经网络逼近任何函数的能力基于众所周知的通用表示定理 (Cybenko, 1989; Hornik, 1991)。该定理表明 $f(x)$ 可以由 $\sum_{j=1}^{\infty} \alpha_j g(w_j^T x + b_j)$ 表示, 其中 g 可以是具有 $\lim_{t \rightarrow -\infty} g(t) = 0$ 和 $\lim_{t \rightarrow +\infty} g(t) = 1$ 的任何连续单调函数。函数 g 被称为激活函数。一些常用的激活函数包括修正线性单元 (ReLU) 函数 $g(x) = \max(0, x)$ 和 sigmoid 函数 $g(x) = 1/(1 + \exp(-x))$ 。在实践中, $f(x)$ 通过有限宽度的神经网络逼近, 即 $\sum_{j=1}^m \alpha_j g(w_j^T x + b_j)$, 其中 $(\alpha_j, w_j, b_j, j=1, 2, \dots, m)$ 是模型参数。

为了很好地逼近一个函数, 单层神经网络的宽度可能必须非常大。深度神经网络通过具有多层, 每层代表一个较短宽度的神经网络, 可以克服这一问题。

一个 L 层的深度神经网络的描述如下。

$$\begin{aligned} \alpha_{1, net} &= \omega_1^T x_i + b_1, \alpha_{1, out} = g_1(\alpha_{1, net}) \\ \alpha_{2, net} &= \omega_2^T \alpha_{1, out} + b_2, \alpha_{2, out} = g_2(\alpha_{2, net}) \\ &\dots \\ \alpha_{L, net} &= \omega_L^T \alpha_{L-1, out} + b_L, \hat{y}_i = \alpha_{L, out} = g_L(\alpha_{L, net}) \end{aligned}$$

其中 x_i 表示输入数据向量。因此, 输出 $\hat{y}_i$ 是参数 $\theta = (w_1, w_2, \dots, w_L, b_1, b_2, \dots, b_L)$ 的复合函数。

在本研究中, 每个订单的销售额是被预测的关键特征, 即该模型的预测目标, 而表一中除销售额外的所有变量都是模型的预测依据, 即预测的特征变量。数据被分为 80% 的训练子集和 20% 的测试子集。在模型归纳之前, 通过应用 “StandardScaler” 实现了特征标准化*, 从而优化了训练过程。

本文所研究的神经网络模型架构由多层组成,

分别配置有 64 个、64 个和 32 个神经元, 并以终端输出节点结束。模型的编译使用了 “adam” 优化器, 以预测结果对实际观察值的均方误差指标作为损失计算的标准。训练协议在 100 个时期的范围内进行, 批量大小为 32。此外, 为了更加详细地监视模型的学习进展, 训练集中包含了 20% 的数据作为验证子集。这一设置有助于在整个训练过程中更准确地评估模型的性能和学习效果。

(三) 结果分析

1. 数据的描述性统计结果。在本文的样本数据中, 含糖饮料在六个城市 22 个区共有 2339 笔订单, 总销售额为 65609.29 元。值得注意的是, 如图 5 所示, 在六个主要城市中, 上海和杭州是通过外卖服务消费含糖饮料最多的两个城市, 其销售额明显高于其他城市, 占总销售额的近 70%。一方面, 这可能是不同的地域饮食文化造成的, 奶茶等各种休闲饮品作为台湾饮食文化的一个代表, 最早的影响是从上海开始, 杭州紧随其后, 而作为精致生活代表的上海和杭州, 对奶茶、咖啡等饮品的消费明显高于其他地区; 另一方面, 也可能是由于长三角地区本身外卖行业的市场份额比重较高, 疫情后外卖行业发展尤为迅速所造成。而含糖饮料作为外卖餐饮中的一个重要部分, 其销售额也明显高于其他外卖发展相对滞后的城市。接下来进一步分析区级层面的销售额情况。

在对这六个城市内各区销售分布进行更详细的分析时, 图 6 呈现出了一个显著的趋势, 杭州拱墅区、上海浦东区和广州番禺区的销售额最高, 杭州西湖区和上海长宁区紧随其后, 这与前面对城市消费分布的分析是趋于一致的。从地理位置来看, 这些区域都位于长三角或以南地区。从区域的经济和人口特征来看, 这些区域也有一些共同点, 包括更高的 GDP、快速发展的外卖行业、繁荣的电子商务、相对数量较多的人口、更为集中的白领人群、偏热的天气、较高的收入水平和快节奏的现代生活方式。

除了考察地理相关特征之外, 本文还调查了通过外卖平台销售含糖饮料的时间模式。如图 7 所

* 这种标准化处理是通过减去特征的均值并除以特征的标准差使得每个特征变量的平均值为 0, 标准差为 1

示,一周内的分布呈现出明显的时间特点。具体而言,与周末相比,工作日的消费量较大,这可能由于大部分人在周末更倾向于在家用餐或外出堂食,而不是外卖。值得注意的是,星期三是一周内的高峰,可能是因为工作日的第三天大部分人感到来自工作或学习的疲劳和压力,而含糖饮料(特别是样本中的奶茶)是一种令人愉悦的茶点,人们往往在繁忙的工作和学习之余订购以放松自己的压力,缓解午后的困倦,当他们没有时间或途径享受其他愉悦事物时,这成为了一种选择,因此周三下午的外卖饮品成为舒缓生活节奏的一种方式。此外,图7右图分析了昼夜模式,发现消费者倾向于在下午下单购买含糖饮料,而在夜间的订单较少,这进一步证实了本文的推论。另外,本文还观察了销售量与

价格之间的关系,图8显示了一个分散的相关性,即较大的订单与平均价格低或高之间并没有太明显的关联。这说明市场行为复杂,可能受到各种外部因素的影响。图8中点的大小代表了销售数量。结果显示,平均价格较低的产品往往有较高的销售数量,但不一定带来更高销售额。然而,某些均价较高的产品同时拥有较高的销售额,也暗示了市场细分可能不太受价格的影响。数据点没有向明确的趋势线收敛,表明单独的价格不会对销售量起决定性影响,暗示着多方面因素会影响消费者的购买决策。

2. 神经网络模型验证。本部分将对线性回归模型与神经网络模型进行比较分析,以验证神经网络模型的有效性。

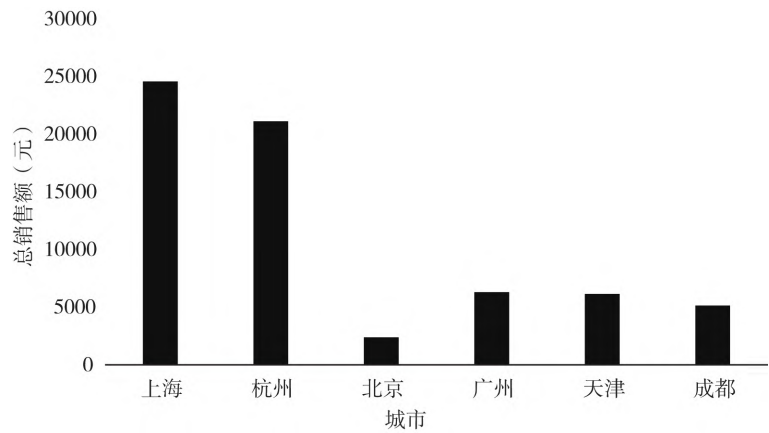


图5 各城市总销售额

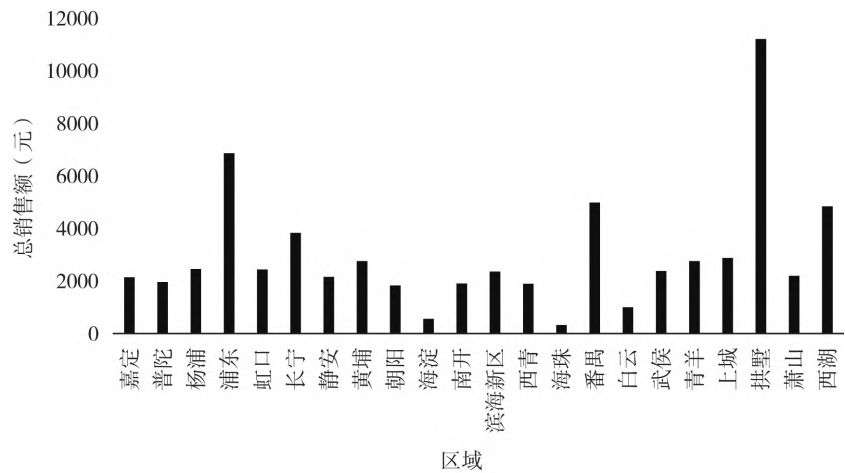


图6 中国各省份的各区域总销售

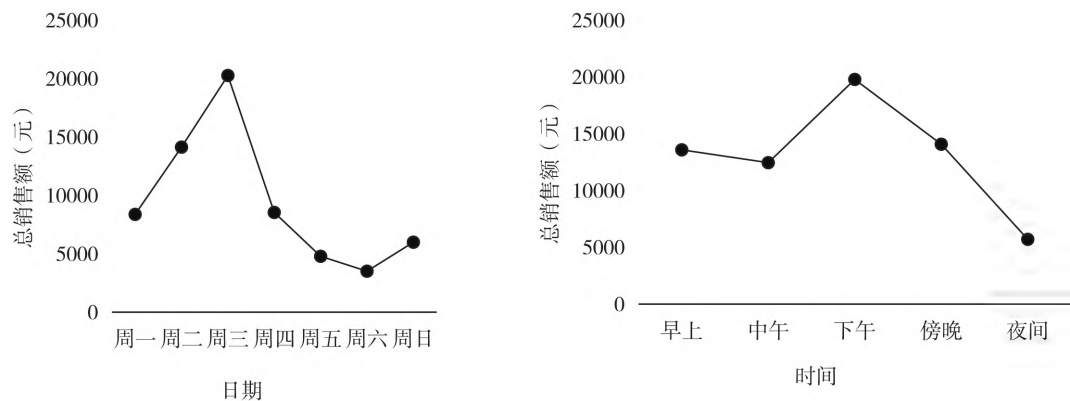


图 7 一周内(左)和一天内(右)的总销售额

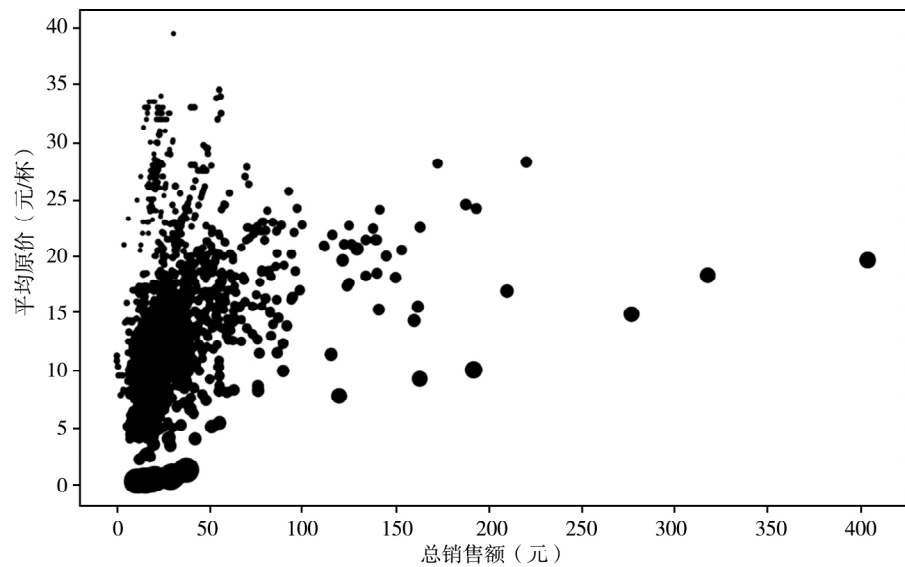


图 8 销售表现分析

首先 构建基准的线性回归模型:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

同样地 模型中 y (被解释变量) 是订单销售额 $x_1, x_2, \cdots, x_n$ 是表一中除销售额外 的变量(解释变量) β_0 是截距项 $\beta_1, \beta_2, \cdots, \beta_n$ 是自变量的系数。 ε 代表误差项 它说明了模型预测值和实际观测值之间的差异。线性回归模型的本质是通过找到确定一组权重 即系数 $\beta_1, \beta_2, \cdots, \beta_n$ 和截距 β_0 , 以最小化误差项 ε ,使得预测的被解释变量值尽可能接近实际观测值。在训练过程中 线性回归模型旨在通过最小化预测值和实际值之间的差的平方和 (MSE ,通常称为均方误差) 来找到最佳拟合直线。

接着 通过对训练轮次内损失函数的收敛性评

估深度学习模型的性能。训练过程显示损失随着训练轮数增加而显著减少 表明模型在预测目标变量方面的准确性在提高。如图 9 所示 ,训练损失和验证损失都迅速下降并趋于收敛 表明模型能够有效地从数据中学习而不过度拟合。在预测方面 ,一百次迭代的水平足以用于模型。由于本文在样本中的可解释变量有限 ,验证模型的损失值将在大约 350 的水平上持续。

最后 ,用预留的 20% 数据来比较基于 80% 数据训练出来的两个模型的预测能力差异。相对于经济学使用的传统线性回归模型 神经网络拟合的模型产生了更好的预测结果。神经网络预测有更低的误差 均方误差分别为 284.90 和 291.83。这个误差主要因为出于隐私保护 数据中没有包含消

费者的任何信息,只有订单数据,造成解释变量的严重缺失。尽管与传统计量回归方法均方误差差异并不是很大,但影响销售因素的机制是非常不同的。况且,一般回归模型要考虑解释变量缺失、内

生等造成的有偏问题,本文的数据正好也存在这样的问题,但神经网络不需要考虑这些问题。下面将应用神经网络模型进行不同的情景模拟,以得出一些经济学的分析结果。

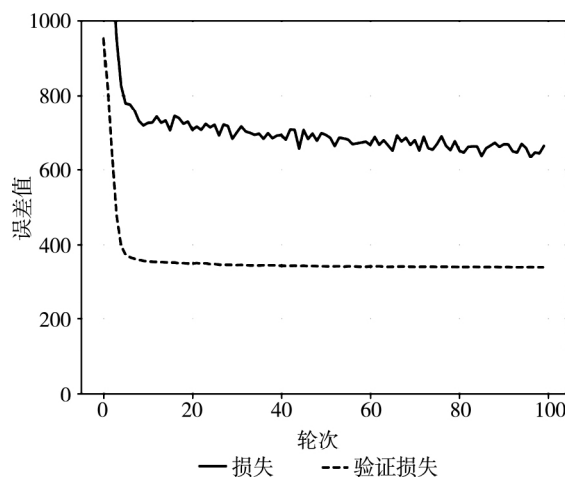


图9 机器学习模型评估

3. 情景模拟分析。本部分将探讨使用预测集数据进行不同情景模拟分析的结果^{*}。首先,基于鲜制含糖饮料店铺本身可以控制的条件变动进行分析,当前数据包含了许多商家提供的促销活动,分析移除所有折扣后,销售将如何受到影响。基于预测样本的模型分析结果显示,在没有折扣的条件下,假设其他条件不变,净销售额几乎没有发生变化。这意味着随着折扣的移除,尽管商家的销售单价变相上涨了,但由于销售量降低,所以净销售额没有受到影响。

考虑到店铺促销活动的复杂性,例如有些店铺在订单达到最低水平时提供固定金额的折扣,而另一些店铺则提供免费配送等服务。为了更深入的分析,设定一个假设情景,即在其他条件不变的情况下,所有商品价格都降低10%。结果显示总销售额相比原始销售额上升了0.49%。这意味着销售数量有所增加,且增加的幅度超过了价格下降的幅度。这表明外卖含糖饮料的需求价格弹性大于1,

即其消费是富有价格弹性的。价格下降能够促进销售额的增加。

接下来考虑糖税的可能影响,由于糖只是含糖饮料的成分之一,所以不能准确估计糖税税率对消费的直接影 响,但可以构建一个类似的假设情景。在此场景中,假设其他条件保持不变,产品价格提高了10%。此时,总销售额将比原始水平下降0.17%,这代表销量减少的幅度超过了价格上涨的幅度,说明价格的增加会显著减少鲜制含糖饮料的消费。进一步分析,如果价格上涨是由征税所引起的,此时征收的税款归政府所有,不应该再计入商家的总销售额,那么商家所得到的总销售额降低幅度应该更大,为10.19%。因此可以推断,如果政府征收糖税,无论是向生产者还是消费者征收,都会导致鲜制含糖饮料价格的上涨,使得其产品销量降低,这表明糖税能够成为减少鲜制含糖饮料消费的可行路径。

四、结论及政策启示

公众已经普遍认识到含糖饮料对健康产生了不良影响。在消费者越来越关注碳酸饮料消费的

同时,一种新型的含糖饮料——奶茶,以在线餐饮服务的形式在城市中迅速成为一种受欢迎的新饮

* 预测结果中的单笔订单预测销售额等于0即为不购买商品,所有订单销售额预测值没有小于0的情况出现

料,尤其受到中国消费者的喜爱。在线餐饮服务订购平台使消费者能够更方便地订购这些含糖饮料。本研究使用了国内一家大型外卖平台的数据和机器学习方法中的神经网络模型,分析了含糖饮料的消费模式。本文的研究是农业和食品经济领域少数应用神经网络方法的研究之一。

研究发现,尽管受制于有限的数据和变量,存在一定的预测误差,但相较于传统的线性回归方法,机器学习方法在本研究中表现出更好的数据拟合效果。实证结果显示,中国城市消费者对价格变化的反应富有弹性。虽然本研究没有消费者数据来证明,但可以侧面反映出这些消费者的年龄层主要是工作年龄或大学年龄的人。他们建立的消费习惯将在很长一段时间内影响他们的生活方式,它对健康的影响可能会在未来几年显现。由于许多国家基于健康问题采取了多项面向含糖饮料行业的政策举措,例如征收糖税,本文研究结果表明征收糖税将显著降低奶茶销量,通过税收手段来减少含糖饮料的消费是一个可行的策略。

本文的研究结果还表明,奶茶、咖啡和各种类型的鲜制含糖饮料主要在工作日下午通过外卖平台订购,作为消费者在一天最繁忙的时候释放压力并与同事和朋友社交的茶点而被消费。它们并不是与正餐一同食用。这意味着从这些含糖饮料中摄入的热量很可能是额外的,而不是来自消费者正常三餐摄入的热量。此外,这种销售不仅局限于中国一线城市,还涉及二线城市。中国南方城市的销售额较高,尤其是长三角地区明显高于其他地区。

五、机器学习在农产品和食物消费研究的未来展望

机器学习可以广泛应用于农产品市场价格波动分析和预测。粮食等重要农产品市场的稳定发展是保障国计民生的关键,2023 年中央“一号文件”强调要抓紧抓好粮食和重要农产品稳产保供。在国际贸易形势复杂多变,新冠、猪瘟等流行疫病频发的背景下,建立实时的机器学习模型以帮助决策者能够及时对市场变化做出反应,对于保障重要农产品市场的稳定发展具有关键意义。有效的预测,可以在调节进出口利用国际市场为保证国内粮食安全争得时间。农产品市场预测影响因素复杂,机器学习相较于传统计量模型能够在应用中从以

这似乎不是源于收入效应,也许是因为温暖的天气使得冷饮更受欢迎,也许是不同地域的饮食文化差异造成的,或者只是这个行业在长三角地区开始并在那里发展得更早。

基于上述结果,本文提出以下政策建议。首先,鉴于糖税在降低奶茶销量上的有效性,可以考虑实施糖税政策,以减少奶茶等含糖饮料的销售。直接给糖加税比加在饮料上更容易操作,可以根据具体情况确定税率和税收水平,以平衡税收收入和健康效益。进一步地,可以考虑逐渐增加含糖饮料的税收,确保生产商和现制饮品店家逐渐减少饮料中的糖分含量,通过渐进的税收制度激励高糖饮料的价格更贴近实际社会成本。其次,考虑到消费者鲜制含糖饮料主要在工作日下午通过外卖平台订购,很可能是正餐外的额外摄入。政府和健康机构应该加强健康饮食的宣传和教育,减少糖分的过度摄入。可以通过学校、社区和媒体传播含糖饮料与肥胖、糖尿病等疾病之间的关联,提高公众对健康饮食的认知,培养健康的饮食习惯。促进健康饮食环境的形成,降低社会对含糖饮料的依赖,从而减少与之相关的健康问题,提高社会的整体健康水平。最后,考虑到奶茶销售在区域和时间上的特点,政府可以根据地域特点制定不同的政策措施,可以选择长三角地区作为引导含糖饮料过度消费的政策先行地区。

在本文研究的基础上,未来研究可以结合更加细致的订单数据或消费者调研数据,识别含糖饮料价格等特征变化对于不同人群消费的影响。

下几点进行突破,首先,机器学习的优势之一是能够处理多源、异构的数据。在农产品市场预测中,除了传统的结构性数据(如产量、库存、需求等)外,机器学习中的自然语言处理技术可以用于对新闻、政策文本等非结构数据进行情感分析、主题建模等,有助于理解市场参与者的态度和看法,提供更深入的市场洞察,从而提供更全面的信息。这有助于更准确地捕捉市场的动态变化和不确定性因素。其次,机器学习可以有效应用图像数据,如遥感数据,用于监测农田状况、作物生长情况、灾害影响等,从而实现更精准的农业监测和预测。最后,

相对于传统计量方法,机器学习在面对高频及复杂变量的数据时可以利用无监督等降维方法提取关键特征,更好地处理不相关变量,更好地捕捉数据的模式和趋势以达到更加精准地预测效果。

另一方面,机器学习在食品安全监测与管理分析中也具有广阔的应用前景。民以食为天、食以安为先。食品安全是保障民生的基础工程,一直以来得到了政府与学者的高度重视。在当前信息化时代,互联网已经成为了食品安全事件与信息分享和讨论的主要渠道,包括政府公文发布、用户购买后的评论以及社交媒体上的信息分享等。传统依赖抽检为主的食品安全监测系统在应对快速传播的大规模信息时显得应对不足。在这一背景下,通过机器学习技术,如网络文本爬取和自然语言处理,可以系统地收集和分析互联网上的食品安全信息。这种方法不仅提高了监测的时效性,还能更全面地了解食品安全事件的发展趋势。此外,建立食品安全监测与预警体系,有助于更迅速地响应食品安全问题,减少潜在风险。并且,机器学习还能够辨识和分析大规模的数据,发现潜在的食品安全问题,有助于在问题蔓延之前及时采取有效的管理和控制措施。因此,机器学习在食品安全监测与管理中的应用不仅能够提高效率,同时也能强化对公众食品安

全的保障。

此外,食物获取和营养摄入与健康息息相关。市场销售数据和人口健康数据在许多国家、省、县市等行政单位都有实际观察记录。虽然这些数据不是抽样调研数据没有个体层面的连接,但区域性数据通常拥有大样本量和丰富的变量,利用机器学习分析这些数据可以对健康问题提供更好的预测,给国家卫生健康部门在医疗资源分配计划中提供更精准的政策建议。

最后,需要注意的是,机器学习在农业经济研究应用需要政策的配套和支持。机器学习的优势在大数据支撑下更为明显,大科研工作人员得以获取到更多的观察值、更多的变量,然而它们往往难以通过现有的计量经济学方法在没有任何信息损失的情况下处理和利用。机器学习的出现改变了这一现状。然而,机器学习的这一优势在样本量和变量都不大的情况下难以充分发挥。为了让机器学习在中国农业经济研究中得到更加充分的利用,研究人员必须获得可用的数据,这要求社会提供更宽松的数据资源共享环境。例如,政府主导的统计数据、或者政府资助的研究机构采集的数据可以考虑在国家信息安全保障的前提下给科研机构免费使用。

参考文献

1. Alzubaidi J., Zhang J., Humaidi A. J., Al-Dujaili A., Duan Y., Al-Shamma O. J., Santamaría A., Fadhel M., Al-Amidie M., Farhan L. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *Journal of Big Data* 2021(8) : 1~74
2. Athey S. The Impact of Machine Learning on Economics. *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, 2018: 507~547
3. Bell J. *Machine Learning: Hands-on for Developers and Technical Professionals*. John Wiley & Sons, 2020
4. Belloni A., Chernozhukov V., Hansen C. High-dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives* 2014(2) : 29~50
5. Bengio Y. Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning* 2009(1) : 1~127
6. Brynjolfsson E., McAfee A. *The Second Machine Age: Work, Progress, and Prosperity in A Time of Brilliant Technologies*. WW Norton & Company, 2014
7. Chen S. H. *Big Data in Computational Social Science and Humanities*. Springer, 2018
8. Cho M., Bonn M. A., Li J. J. Differences in Perceptions about Food Delivery Apps between Single-person and Multi-person Households. *International Journal of Hospitality Management* 2019, 77: 108~116
9. Cybenko G. Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals and Systems*, 1989(4) : 303~314
10. Dai X., Wu L., Hu W. Nutritional Quality and Consumer Health Perception of Online Delivery Food in the Context of China. *BMC Public Health* 2022(1) : 1~12
11. De Freitas R S G., Stedefeldt E. COVID-19 Pandemic Underlines the Need to Build Resilience in Commercial Restaurants' Food Safety. *Food Research International* 2020, 136: 109472
12. Gao Y., Zheng Z., Henneberry S. R. Is Nutritional Status Associated with Income Growth? Evidence from Chinese Adults. *China Agricultural*

- Economic Review 2020(3) : 507~525
13. Gopinath , M. , Batarseh , F. A. , Bechman , J. , Kulkarni , A. , Jeong , S. International Agricultural Trade Forecasting Using Machine Learning. Data & Policy 2021 3: e1
 14. Graskemper , V. , Yu , X. , Feil , J. H. Farmer Typology and Implications for Policy Design—An Unsupervised Machine Learning Approach. Land Use Policy 2021 ,103: 105328
 15. Gupta , M. A Study on Impact of Online Food Delivery App on Restaurant Business Special Reference to Zomato and Swiggy. International Journal of Research and Analytical Reviews 2019(1) : 889~893
 16. Hastie , T. , Tibshirani , R. , Friedman , J. H. et al. The Elements of Statistical Learning: Data Mining ,Inference and Prediction. Springer 2009
 17. Hodeghatta , U. R. , Nayak , U. Unsupervised Machine Learning. In: Business Analytics Using R – A Practical Approach. Apress ,Berkeley , CA 2017
 18. Hornik , K. Approximation Capabilities of Multilayer Feedforward Networks. Neural Networks ,1991(2) : 251~257
 19. Hu , F. B. , Malik , V. S. Sugar – sweetened Beverages and Risk of Obesity and Type 2 Diabetes: Epidemiologic Evidence. Physiology & Behavior 2010(1) : 47~54
 20. Jordan , M. I. , Mitchell , T. M. Machine Learning: Trends ,Perspectives and Prospects. Science 2015(6245) : 255~260
 21. Kotsiantis , S. B. Supervised Machine Learning: A Review of Classification Techniques. Informatica 2007 31: 249~268
 22. Kumar , S. , Shah , A. Revisiting Food Delivery Apps During COVID–19 Pandemic? Investigating the Role of Emotions. Journal of Retailing and Consumer Services 2021 62: 102595
 23. Lan , H. , Li , Y. , Wang , H. Improvement of Online Food Delivery Service Based on Consumers’ Negative Comments. Canadian Social Science , 2016(5) : 84~88
 24. Liu , C. , Zhou , L. , Höschle , L. et al. Food Price Dynamics and Regional Clusters: Machine Learning Analysis of Egg Prices in China. China Agricultural Economic Review 2023(2) : 416~432
 25. Lobstein , T. , Brinsden , H. Atlas of Childhood Obesity. World Obesity Federation 2019
 26. Maimaiti , M. , Zhao , X. , Jia , M. et al. How We Eat Determines What We Become: Opportunities and Challenges Brought by Food Delivery Industry in a Changing World in China. European Journal of Clinical Nutrition 2018(9) : 1282~1286
 27. Mason , W. , Vaughan , J. W. , Wallach , H. Computational Social Science and Social Computing. Machine Learning 2014 95: 257~260
 28. NBS. 2021 China Chain Catering Industry Report. National Bureau of Statistics ,China Internet Network Information Center(CNNIC) 2022
 29. Popkin , B. M. , Hawkes , C. Sweetening of the Global Diet ,Particularly Beverages: Patterns ,Trends and Policy Responses. The Lancet Diabetes & Endocrinology 2016(2) : 174~186
 30. Roh , M. , Park , K. Adoption of O2O Food Delivery Services in South Korea: The Moderating Role of Moral Obligation in Meal Preparation. International Journal of Information Management 2019 47: 262~273
 31. Ruiz , F. J. , Athey , S. , Blei , D. M. SHOPPER: A Probabilistic Model of Consumer Choice with Substitutes and Complements. 2020
 32. Schmidhuber , J. Deep Learning in Neural Networks: An Overview. Neural Networks 2015 61: 85~117
 33. Shankar , A. , Jebarajakirthy , C. , Nayal , P. , et al. Online Food Delivery: A Systematic Synthesis of Literature and A Framework Development. International Journal of Hospitality Management 2022 ,104: 103240
 34. Singh , G. M. , Micha , R. , Khatibzadeh , S. , et al. Global Burden of Diseases Nutrition and Chronic Diseases Expert Group(NutriCoDE) . Global Regional and National Consumption of Sugar–Sweetened Beverages ,Fruit Juices and Milk: A Systematic Assessment of Beverage Intake in 187 Countries. PLoS ONE 2015 ,10 e0124845
 35. Statista. eServices Report 2020—Online Food Delivery <https://www.statista.com/study/40457/foo>
 36. Stephens , J. , Miller , H. , Miletello , L. Food Delivery Apps and the Negative Health Impacts for Americans. Frontiers in Nutrition 2020(7) : 14
 37. Storm , H. , Baylis , K. , Heckelei , T. Machine Learning in Agricultural and Applied Economics. European Review of Agricultural Economics , 2020(3) : 849~892
 38. Suhartanto , D. , Helmi Ali , M. , Tan , K. H. et al. Loyalty Toward Online Food Delivery Service: The Role of E–Service Quality and Food Quality. Journal of Food Service Business Research 2019(1) : 81~97
 39. Taddy , M. The Technological Elements of Artificial Intelligence ,In: A. K. Agrawal et al(eds) . Economics of Artificial Intelligence ,University of Chicago Press 2018: 61~87
 40. Varian , H. R. Big Data: New Tricks for Econometrics ,Journal of Economic Perspectives. Journal of Economic Perspectives 2014(2) : 3~28
 41. World Health Organization. Global Prevalence and Secular Trends in Obesity. In “Obesity Preventing and Managing the Global Epidemic ,Report of a WHO Consultation on Obesity. ” Geneva: WHO ,1998: 17~40
 42. World Health Organization. Guideline: Sugar Intake for Adults and Children , Geneva: WHO 2015

43. Yeo ,V. C. S. ,Goh ,S. K. ,Rezaei ,S. Consumer Experiences ,Attitude and Behavioral Intention Toward Online Food Delivery(OFD) Services. *Journal of Retailing and Consumer Services* ,2017 ,35: 150~162
44. Zeng ,Q. ,Chen ,X. Identification of Urban – Rural Integration Types in China – an Unsupervised Machine Learning Approach. *China Agricultural Economic Review* ,2023(2) : 400~415
45. Zhang S. ,Luan ,H. ,Zhen F. ,et al. Does Online Food Delivery Improve the Equity of Food Accessibility? A Case Study of Nanjing ,China. *Journal of Transport Geography* ,2023 ,107: 103516
46. Zhao ,H. Futures Price Prediction of Agricultural Products Based on Machine Learning. *Neural Computing and Applications* ,2021 ,33: 837~850
47. Zhou ,Y. ,Shan ,Y. ,Guan ,D. ,Liang ,X. ,Cai ,Y. ,Liu ,J. ,Xie ,W. ,Xue ,J. ,Ma ,Z. ,Yang ,Z. Sharing Tableware Reduces Waste Generation , Emissions and Water Consumption in China’ s Takeaway Packaging Waste Dilemma. *Nature Food* ,2020(9) : 552~561
48. 樊胜根 ,张玉梅. 践行大食物观促进全民营养健康和可持续发展的战略选择. *农业经济问题* ,2023(5) : 11~21
49. 全世文. 肥胖脆弱性: 概念与测度. *中国农村观察* ,2021(6) : 2~25
50. 陶旭辉 ,郭 峰. 异质性政策效应评估与机器学习方法: 研究进展与未来方向. *管理世界* ,2023(11) : 216~237
51. 王 川 ,王 克. 基于 BP 神经网络的我国农产品市场风险预警研究. *农业经济问题* ,2008(S1) : 152~156
52. 于晓华 ,唐 忠 ,包 特. 机器学习和农业政策研究范式的革新. *农业技术经济* ,2019(2) : 4~9

A Consumption Analysis of Sugar-sweetened Beverages Ordered Online for Home Delivery: An Application of Machine Learning

ZHAI Qianqian ,JIANG Yu ,ZHANG Jiaming ,Juan Esteban GUEVARA ,WANG Hong

Abstract: Sugar-sweetened beverages have caused worldwide health concerns due to their contribution to high calorie intake and chronicle diseases. Against the backdrop of the rapid development of online food delivery platforms ,this paper utilizes sales data from delivery services in six Chinese cities and employs machine learning methods to identify consumption patterns of sugar-sweetened beverages. Additionally ,the paper analyzes sales variations in simulated scenarios such as price discounts and sugar taxes. The findings of the paper are threefold: First ,neural networks within supervised machine learning models demonstrate greater applicability for predicting consumption trends compared to traditional econometric regression models. Second ,model predictions indicate that a 10% price reduction leads to a 0.49% increase in total sales ,suggesting significant price elasticity of demand for sugar-sweetened beverages. Conversely ,a 10% price increase results in a 0.17% decrease in total sales volume ,implying that the utilization of taxation as a strategy to reduce the consumption of these beverages is viable. Third ,there are spatial and temporal variations in the sales of sugar-sweetened beverages on delivery platforms ,with higher sales volumes in the Yangtze River Delta region compared to other areas , and weekday sales surpassing those of weekends ,with afternoons being the peak period. This research is significant for establishing pathways and mechanisms to reduce excessive consumption of sugar-sweetened beverages ,guiding consumers towards healthier dietary choices ,and promoting sustainable development in the food delivery service industry.

Keywords: Sugar-sweetened beverages; Delivery; Machine learning; Neural networks

责任编辑: 吕新业