

大数据剖析生鲜电商平台销量的影响因素

——基于支持向量机方法的实证研究

林海 胡雅淇¹

(中国农业大学, 北京 100083)

【摘要】: 电商能够减少农产品流通环节, 促进小农户与大市场的对接。利用 Python 爬虫工具, 以羊肉产品为例, 采集淘宝平台的生鲜电商数据, 中小型生鲜电商的特征指标、区域分布以及经营规模显示淘宝平台的生鲜电商存在现有规模较小、区域间差异较大、产品以整斤包装为主、坏单包赔等交易保障有限等特征。基于支持向量机(SVM)的方法, 根据选取的特征指标, 设定不同方案对羊肉产品销量进行预测, 判断中小型生鲜电商销量的影响因素, 结果表明: 评论特征、交易特征、店铺特征、服务特征以及产品特征对产品销量均会产生程度不同的影响, 其中以评论特征的影响最为显著。基于此, 平台上的卖家要以销定产, 制定更加合理的包装规格和单品价格并提供适当的保障措施。政府应根据地区不同优势, 进行政策引导与支持, 制定网销生鲜产品分类标准等, 降低消费者购买风险。电商平台优化在线评论界面, 提高规范化店铺入驻管理, 减少信息不对称。

【关键词】: 生鲜电商 大数据 销量分析 支持向量机(SVM)

【中图分类号】: F325. 2 **【文献标识码】:** A **【文章编号】:** 1002-6924(2021)3-0129-10

一、引言

如何实现小农户和大市场的有效对接, 一直是保障农产品有效供给, 促进农民收入持续稳定增长的关键问题之一, 在小农户长期存在的背景下, 解决该问题是乡村振兴面临的重大课题。随着信息技术的持续发展和新生代消费者的崛起, 电子商务的触角已经由工业品伸向农产品, 与传统的销售方式相比, 电子商务能够结合信息沟通快捷的互联网行业、发展迅猛的快递行业和满足多样消费需求的网络电商平台三方面的优势, 减少产品流通的交易环节, 提高农业产业链整合度, 有效降低农产品市场的交易成本。在消费不断升级、生鲜产品市场需求日益增长的背景下, 生鲜电商作为农户进入市场的新兴渠道, 一方面, 能够方便个体农户与大市场的有效对接, 从而一定程度上促进农民持续增收; 另一方面, 有助于提高企业工业化和信息化的融合程度, 增加消费者效用水平。

当前, 生鲜电商的发展既受到政策的支持以及企业资本的关注, 又受到产品特性及技术条件的限制, 在二者双重作用下, 生鲜电商不断向前发展。政策层面上, 国家已经开始高度重视农产品电商在发展现代农业、提高农民福利中的作用。2019 年中央一号文件指出: “继续开展电子商务进农村综合示范, 实施‘互联网+’农产品出村进城工程”是实施数字乡村战略的重要手段和措施, 并强调“扶持小农户和现代农业发展有机衔接”。此外, 生鲜电商顺应了消费升级的需求, 不断带来产业新业态、商业新模式以及经济新增长点, 因而备受资本市场的重视, 随着阿里、京东等电商巨头不断加大在冷链物流和生鲜供应链方面

¹**作者简介:** 林海, 中国农业大学经济管理学院副教授, 博士生导师, 主要研究方向: 农产品贸易、农村发展;

胡雅淇, 中国农业大学经济管理学院博士研究生, 主要研究方向: 数字农业、农产品电商。

基金项目: 农业农村部“国家现代肉羊产业技术体系”(CARS-38)

的投资,创新模式层出不穷,生鲜电商市场进入新的格局。此外,受限于生鲜产品的特性及现有技术条件,生鲜电商的发展也受到诸多因素的限制,例如,生鲜产品易腐烂、难储存的缺陷导致的巨大的物流成本;冷藏运输技术等方面存在滞后和不足;消费者亲身体验获取商品信息的能力弱化造成的信息不对称问题等。

无论是理论层面还是政策层面,生鲜电商的发展现状和驱动及限制因素都备受关注。在此形势下,代表性电商平台上的中小型生鲜电商经营状况如何(规模、产品特征、区域差异)?哪些因素影响了其网络销售行为?这些因素的影响程度究竟如何?目前仍需要更加全面的数据及进一步研究。本研究旨在利用 Python 爬虫采集的大数据“2018 年淘宝羊肉电商评论数据集”,基于机器学习前沿方法,即支持向量机(SVM),来剖析影响产品销量的因素,从多维度探究生鲜电商的发展状况。从方法上,采用不同于传统计量方法的大数据方法,因而在数据的采集、清洗、分析、使用等方面均体现出新的特征,^[1]大数据与生鲜电商结合的研究一方面可以利用实时数据,对生鲜电商的现状进行更为直观的定量分析;另一方面,基于经济学理论的相关性分析,能够发现数据间的联系,探索生鲜电商发展的影响因素。从产品上,以羊肉产品为例,进行生鲜电商销量影响因素的分析,一是因为羊肉产品是高蛋白、低脂肪类的健康食品,提高此类食品的消费比例是中国未来居民食物结构调整的重要方向,^[2]是消费升级、消费者福利增长的重要体现;二是因为羊肉生产规模化程度较低,小规模生产所导致的产销衔接矛盾突出,在此情况下,电商平台对于小农户对接大市场的意义就更为重大;三是因为与其他生鲜产品相比,肉类产品的检疫更为复杂,电商销售面临更多产品质量的不确定性问题,对它的分析更能反映出信息不确定环境下的交易特征。

鉴于生鲜电商在市场规模和用户规模上都存在着巨大的增长空间,全面剖析其发展现状,深度挖掘其驱动和限制因素,对于掌握“互联网+农业”产业重塑存在的问题、把握产业链结构升级的方向、更好地利用生鲜电商促进小农户与大市场的有效对接均具有重要的现实意义。本文创新之处在于使用了 Python 爬虫获取的大样本数据,将机器学习 SVM 的方法与生鲜农产品电商的研究相结合,围绕电商综合性平台的公开指标,比较系统地回答了上述问题,并对不同利益主体提出有针对性的政策建议。

二、文献评述

随着我国数字经济的深入发展,农产品电子商务的相关研究也在不断增加,所采用的方法也持续更新。在研究内容上,首先,部分研究集中在企业电商采纳的行为分析上,包括技术、服务、组织等在内的因素都会影响企业电商的运营,但研究涉及的因素较多、未形成统一的研究体系。Canavari et al.^[3]指出 B2B 平台的信任问题是农产品电商采纳率很低的重要原因之一。侯振兴^[4]通过访谈认为基础设施、网络安全、第三方电商服务等多种因素都会影响电商的采纳。林家宝等^[5]认为组织敏捷性对农产品电子商务能力和企业绩效之间的关系存在中介作用。葛继红等^[6]通过研究指出我国生鲜电商的三种主流发展模式分别为“平台+商家”生态圈模式、垂直一体化模式和 O2O 模式,它们所具备的优势不同、适用条件不同。此外,电商的采纳主体研究已由企业逐步涉及到农户,例如,曾亿武等^[7]认为电子商务采纳对农户诚信经营意识有着显著的积极影响,且电子商务采纳对受教育程度较低、品牌化意识较弱以及网店经营期限较长的农户的诚信经营意识具有更强的促进效应。其次,较多文献关注消费者在线购买农产品的行为分析,从购买意愿等视角分析生鲜电商销售的影响因素。Ernst and Hooker 指出消费者认为最重要的影响因素是农产品质量的不确定性,因而农产品质量信号的研究是当前农产品电商领域的研究重点之一。^[8]王克喜和戴安娜^[9]指出农产品中的质量、安全、新鲜度对网购生鲜农产品意愿具有显著正向影响。梁文卓等^[10]认为储存环境要求较低、消费时效较长、物流包装无特殊技术要求、消费风险较低的农产品将成为网络销售的主体。吴自强^[11]通过分析发现,产品种类认知、品牌认知、食品安全与健康支付意愿等因素对消费者生鲜农产品的网购意愿有显著的正向影响。可以看出,当前生鲜电商采纳多以企业为研究主体,对于生鲜电商影响因素的分析大部分是从消费者视角出发,涉及因素维度较广,不同研究设定变量并不统一。

在研究方法上,农业经济领域已开始尝试大数据方法的使用,Cavallo et al.^[12]通过抓取 900 多个零售商的网站上 1500 多万种商品的在线价格,计算每日价格指数作为美国等 20 多个国家每日更新的通货膨胀指数,文章指出由于数据的可广泛获取性,在美国的食品部门和电子部门,在线数据和消费者价格指数匹配程度更高。雷怀英等^[13]以 2011 年至 2017 年为跨度利用网络爬虫技术获取网络检索数据,对网络检索数据与物价波动关系进行分析研究。大数据在农产品消费行为研究的应用是该领域的重要方向,主要集中在消费者在线购买影响因素的分析,通过研究评论特征以及网络口碑等电商大数据,讨论其对消费者决策的

影响。早期研究认为网络口碑能够显著影响到消费者的购买行为和电商平台的销量，^{[14][15]}这种说法得到殷国鹏^[16]、龚诗阳等^[17]研究进一步证实，这些研究的差异主要是在于对具体的网络口碑指标，比如评论数量和评论内容等存在着不同的解释。近期，研究则更进一步，开始考虑网络口碑具体的作用路径和机制^[18]及其在新产品推广中的应用。^[19]王英等^[20]利用网页爬虫技术根据京东的稻花香所有系列产品的为期一年的在线评论数据，探索品牌、在线评论与农产品销量之间的关系以及二者如何对农产品销量产生影响。可以看出，大数据方法在获取过程中，能够对研究对象的数据进行内容更为全面、样本更为普及的搜集，能够有效探索指标之间的相关性，并通过补充样本量、丰富经济指标覆盖维度为实证研究提供更多的算法。

总体上，对于生鲜电商的研究多以企业的采纳行为和消费者的在线消费行为分析为主，而以农户以及个体户为主的中小型生鲜农产品销售为主体的线上平台行为分析研究较少。其中重要的原因之一在于生鲜农产品线上数据的获取有一定困难，且与大数据等前沿分析方法的实践结合还尚待开发完善。目前农产品电商研究现状的分析主要是基于传统的抽样调查方法采集的数据，基于经济学逻辑框架，农业经济领域的大数据研究相对有限。笔者认为，以生鲜电商平台为研究对象，运用大数据的数据抓取方法和分析方法相结合，采集并剖析淘宝平台特定生鲜电商的相关数据，既可以丰富生鲜电商的采纳主体研究维度，有助于统一不同研究之间选取变量的框架，又补充了生鲜电商大数据方法的实践研究，进而完善生鲜电商的经济学分析文献、有利于全面地掌握生鲜电商市场的总体情况和规律。

三、数据来源及研究方法

数据来源于笔者所在团队采集的“2018 年淘宝羊肉电商评论数据集”，该数据来源于中国最大的移动商务平台淘宝网(www.taobao.com)。该研究对数据进行挖掘和分析的方法主要包括两种：(1)运用 Python 爬虫工具进行数据的采集；(2)采用 R 中的 SVM 对产品销量的影响因素进行分类判定，以下对两种方法的流程及具体设计方案进行逐一介绍。

(一)基于 Python 语言的网络爬虫工具

基于 Python 语言的网络爬虫，是根据网页链接地址来查找网页内容，并提取特定信息，将数据返回给用户的网络信息采集方法。^{[21][22]}在互联网大数据的背景下，随着在线购物交易额的迅速增长和市场比例的逐渐上升，通过网络大数据挖掘，分析在线交易特征，将为在线交易平台的发展提供更多有效信息，使网络购物双方都能受益。目前，网络爬虫在经济研究的应用方兴未艾，较多的文献集中在日用消费品、服务等在线评论样本的抓取，利用爬虫方法研究农产品电商领域的文献相对缺乏。孙易冰等^[23]参照官方 CPI 的制度方法，基于网络爬虫技术的价格指数计算模型，发现该种模型具有时效性强和灵敏度高的特点。阮素梅等^[24]通过网络爬虫从人人贷网站上抓取数据样本，对 P2P 借贷中借款人的违约风险进行评估；Ye et al.^[25]通过 Ruby 爬虫从携程网上抓取评论数据样本，研究了在线评论对酒店房间订购量的影响。Ma et al.^[26]利用 Python 和 Java 编写的爬虫程度从 Trip Advisor 和 Yelp 网站上抓取样本，从深度学习的角度分析照片对酒店订购量的影响。霍红等^[27]通过 Scrapy 框架自主编写网络爬虫数据，采集当当网为期三个月的电子产品、数码产品和日化品的评论信息，研究在线评论质量丰富度对功能性产品和享受性产品的不同影响。

(二)定量分析数据获取流程

基于 Python 爬虫，描述统计部分的数据获取及处理过程主要分为以下四个主要步骤：

1. 数据集的具体获取过程。

采用 Python 爬虫工具，利用爬虫中内嵌的关键词匹配模块，根据关键词进行匹配筛选，本数据集以“新鲜羊肉”作为关键词，对淘宝上所有实时销售的羊肉产品进行数据抓取，抓取内容涵盖羊肉店铺相关信息及其产品评论信息。根据程序设定，整个抓取过程在两小时内一次性完成，既提高了效率、节省了内存资源，亦保证了各店铺数据的可比性。

2. 数据清洗、剔除无关数据。

在文本抓取过程中,会出现很多与羊肉相关但不是羊肉的产品,在程序设定中尽量将其剔除,即包含“羊肉汤、羊杂汤、火锅底(调)料”等 86 个词汇的样本,剔除无关样本之后,形成数据集。团队于 2018 年 1 月进行一次性抓取,形成了数据集,其中包括 30316 个含评论的样本。根据研究目的,将样本进行筛选归类,首先按照店铺进行分类,将由于评论指标而出现的重复样本删除,最后产生有效样本,共包含 765 家店铺的 1995 条羊肉产品信息(下文将其视为 1995 种羊肉产品)。

3. 变量的选择和数据格式的转换。

首先,在样本筛选完毕的基础上,挑选出需要的变量:(1)评论特征。包括评论次数(累计评论数量)、评论长度(第一页评论的字数平均值)、好评数量、差评数量、图片数量;(2)交易特征。包括快递费用(包邮=1,不包邮=0)、质量保障(有=1,无=0)、坏单包赔(有=1,无=0);(3)店铺特征。包括担保资金、信誉、所在区域(省份);(4)服务特征。包括产品与描述符合程度、物流水平、服务水平;(5)产品特征。包括产品单价、包装规格(单品重量)、销量(以付款人数为指标)。其次,由于 Python 抓取并清洗过的数据表是以 csv 形式存在的样本,所有信息都在一个单元格内,运用 Excel 数据清理的方法,将数据无关的符号去除,并将数据按照不同变量分割成可处理的形式;最后,将数据导入 Stata,形成易于处理的 dta 格式。

4. 统计分析。

利用 Stata 15 统计软件,依次对包含包装规格、物流水平等在内的生鲜电商的特征指标、生鲜电商区域分布及其经营规模进行描述统计分析。

(三)支持向量机(SVM)的分类器方法

支持向量机(Support vector machine,SVM)最早由 Cortes and Vapnik^[28]提出,是建立在统计学理论上的一种机器学习方法。^[29]目前,支持向量机和随机森林是最准确的分类工具,分类效果明显高于其他众多分类器。^[30]然而,国内将 SVM 分类器的方法应用到经济学领域的研究尚且缺乏。

SVM 可以想象成一个平面,该平面创建了数据点之间的边界,而这些数据点绘制在代表案例及其特征值的多维空间中。^[31]SVM 可以用以下四点简单概括:一是分类:基本上是通过最大化两类数据最接近点间的间隔来寻找最佳分离超平面(参见图 1)——位于边界上的点称为支持向量,而边界的中间部分为最优分离超平面;二是有重叠的分类:有的数据点在判定间隔的“错误侧”会被赋予权重以降低影响,这就形成了所谓的软边界(soft margin);三是非线性可分的情况:当找不到线性分类器时,数据点通常会被投影到高维空间中,通过核函数处理技巧,数据点在高维空间是有效的线性可分;四是解决方法的本质:实际上,整个任务为一个二次优化问题,因此可以通过已知的技术来解决。^[32]

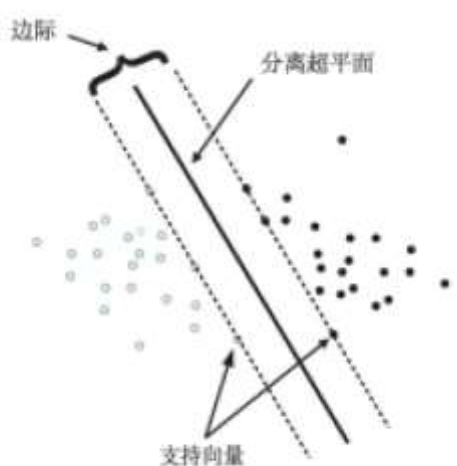


图 1 支持向量机原理图(以线性为例)

（四）SVM 分类的实现方案

SVM 的分类功能通过 R 软件中的“e1071”程序包(package)实现，值得注意的是:在 R 中操作时，要确保被预测变量为因子形式(factor)，这时候实现的才是 SVM 分类器功能，实现流程主要包括以下四个步骤：

1. 选定可能影响销量的变量作为特征指标。

影响生鲜产品销量的因素包含评论特征、交易特征、店铺特征、服务特征和产品特征五大类，具体的变量指标如表 1 所示：

表 1 影响生鲜销量的变量及特征指标

影响变量	特征指标
评论特征	评论次数、评论长度、好评数量、差评数量、图片数量
交易特征	快递费用、质量保障、坏单包赔
店铺特征	担保资金、信誉、所在区域
服务特征	产品与描述符合程度、物流水平、服务水平
产品特征	产品价格、包装规格

2. 将销量数据进行分级处理并标准化处理特征指标。

采纳淘宝平台的付款人数作为销量的代理指标，且假定每次付款只买一件产品，该假定是基于淘宝平台是零售平台而不是批发平台且消费者对生鲜产品的单次消费量并不是很大，为简化研究，因而特作如此假定。首先对销量进行等级划分，将销量分成 8 个等级，如表 2 所示：

表 2 影响生鲜销量的变量及特征指标

销量范围	等级	销量范围	等级
0—10	1	100—200	5
10—30	2	200—500	6
30—50	3	500—800	7
50—100	4	800 以上	8

其次，采用 Z-score 的方法对除销量外的其他指标进行归一化处理，从而标准化原始指标数据。

3. 将数据分为训练集和测试集，并基于训练集训练 SVM 分类器。

首先，将数据按照指标逐一剔除含有空值的样本之后，共有有效样本 1995 个；其次，将样本分为训练集和测试集。参照机器学习方法文献，可以将 3/4 的样本作为训练集，剩下的 1/4 作为测试集，^[33]即训练集的样本数量为 1496，测试集的样本量为 499；也可以将 4/5 的样本作为训练集，剩下的 1/5 作为测试集，^[34]本文采用上述两种样本拆分方法。

4. 基于测试集进行预测以评估模型性能并求出准确率。

基于训练集，采用 SVM 命令进行模型的构建，然后用测试集的数据进行预测，从而判断依据特征指标对销量进行预测的准确度。

5. 尝试不同特征指标并比较准确度。

依次去掉影响销量的五类特征指标的一种，形成五种模拟方案。值得注意的是，SVM 分类器中有四种核函数(kernel):线性(linear)核函数、多项式(polynomial)核函数、S 形(sigmoid)核函数、高斯 RBF (Gaussian RBF)核函数。根据相关研究，高斯 RBF (Gaussian RBF)核函数已被证明对于许多类型的数据都能运行得很好，^[31]因此研究采用基于 RBF 的支持向量机以尽可能的实现模型性能的最优化。同时调整成本约束参数 C(C 越大，margin 越小，允许误差样本越少)来修正决策边界宽度，测试模型估计的准确度。

四、特征指标的定量分析

基于淘宝公开可获取的店铺及经营信息，笔者一方面对羊肉电商店铺特征、服务特征以及产品特征等进行描述性统计，从微观层面剖析个体特征影响因素；另一方面从区域分布及经营规模等方面进行归纳总结，从宏观角度把握羊肉电商现状。基于前文所选定的关键特征指标，笔者的定量分析从各特征指标的平均值、标准差、最小值、最大值、中位数、众数等入手，剖析羊肉电商的现状与特点。

(一)特征指标层面分析

首先，从店铺特征来看，不同店铺的担保资金虽然差距较大，但以 1000 元为主，以 1000 元为担保资金的样本占比为 66.82%；店铺信誉千差万别，高信誉和低信誉店铺之间的信誉差距较大。其次，从服务特征来看，无论描述、物流还是服务水平，各店铺情况相对平均，三个特征指标值以 4.8 为主，描述、物流及服务水平为 4.8 的样本占比依次为 49.77%、45.71%、43.26%。第三，从产品特征来看，包装规格上具有多样性和零碎化的特点，重量从不足一斤到四五十斤各有不等，但主要的包装规格主要为 1—4 斤等整斤包装，且相同重量规格产品的价格差异化程度较高，普遍要高于线下不同市场间的销售价格差异。第四，从交易特征来看，目前羊肉的在线销售提供的服务保障仍较为缺乏，大部分产品都未向消费者提供质量保障和坏单包赔，同时倾向

于收取快递费用。最后，从评论特征来看，评论次数虽然差距很大，但大多数产品的评论次数还是相对有限，而且其中以好评数量为主，差评数量以及上传的图片数量均较少。评论长度相对均匀，处于 1—21 个字之间不等，大多数在 10 字左右。

表 3 羊肉电商特征指标的描述性统计

特征指标	变量名称	平均值	标准差	最小值	最大值	中位数	众数	众数个数
店铺特征	担保资金	204.0	356.0	0	51000	1000	1000	1333
	店铺信誉	32208.9	139634.5	1	2700000	4347	1488	18
服务特征	描述符合	4.8	0.1	3	5	4.8	4.8	993
	物流水平	4.8	0.1	2	5	4.8	4.8	912
	服务水平	4.8	0.1	3	5	4.8	4.8	863
产品特征	产品单价	45.3	22.7	1.7	260	43.6	45	46
	包装规格	4.6	6.7	0.1	50	2	1	661
交易特征	快递费用	0.5	0.5	0	1	0	0	1024
	质量保障	0.3	0.4	0	1	0	0	1464
	坏单包赔	0.4	0.5	0	1	0	0	1140
评论特征	评论次数	87.4	276.4	1	5071	16	1	181
	评论长度	7.0	1.9	1	21	7	9	99
	好评数量	86.3	274.9	0	5069	16	1	179
	差评数量	0.2	0.9	0	23	0	0	1769
	图片数量	4.3	16.0	0	519	1	0	851

注:快递费用(包邮=1,不包邮=0);质量保障(有=1,无=0);坏单包赔(有=1,无=0);担保资金(元);单价(元/斤)

(二)区域分布角度分析

淘宝羊肉店铺在全国 31 个省、市、自治区(不含港、澳、台地区)的分布数量差距较大,省际间发展不平衡、差异较大特征明显。此外,店铺各数量段的情况,反映出对大多数省份而言,羊肉电商的发展仍处于初级阶段。当前大多数省羊肉电商的数量集中在 0—20 个,店铺个数超过 20 个的只有 10 个省份,超过 50 个的只有 5 个省份(见表 4)。

表 4 羊肉淘宝店区域分布情况

店铺数量	省份
100*200	内蒙古
50*100	上海、江苏、山东、甘肃
10*50	广东、宁夏、北京、新疆、浙江
10*30	安徽、四川、陕西、河北、陕西、河南、青海、山西
0*10	湖北、辽宁、福建、黑龙江、重庆、湖南、广西、贵州、江西、海南、云南、天津、西藏、吉林

注:每一店铺数量段的省份排序均是数量由高到低进行,单位(个)

结合电商发展状况、羊肉生产情况及羊肉消费水平,进一步分析羊肉电商发展区域差异可能存在的原因。羊肉淘宝店铺数量前十位中,排名第一的内蒙古羊肉产量最高、消费量第二;排名第二的上海消费量第一;排名第三的江苏是农产品电商发展最好,也是羊肉消费排名第三的;排名第四的山东省羊肉产量第三、羊肉消费第四。总之,淘宝店铺数量前十的省份其一可能是羊肉主产地,其产量在全国范围内具有显著优势;其二可能是农产品电商发展状况良好的省份,其农产品电商环境发展良好;其三可能是羊肉消费水平较高的省份,具有稳定的消费基础。可以看出,生鲜电商的发展或依赖于强大的农业产业优势,或依靠于灵活的电商发展环境,或依托于良好的消费市场。

(三)经营规模角度分析

淘宝的羊肉店铺规模普遍较小,笔者以销量作为衡量店铺经营规模的指标。目前大多数淘宝羊肉店铺的月度成交量都较低,普遍在100单以下,且大部分集中在30单以下。销量在10单以内的产品占比高达67%以上,大多数淘宝羊肉店铺的销量保持在很低的水平,只有个别的店铺产品销量在200单以上。月度总销量以万单为衡量量级,该样本显示约为销量超过4万单。通过经营规模的分析,可以看出淘宝平台的羊肉店铺销量有限,一方面的原因在于淘宝商家以个体户和小企业为主,另一方面也反映出羊肉电商发展规模仍处于较低水平,羊肉电商可发展的空间还是相对较大。

表5 羊肉电商区域分布情况

省份	店铺数量	所占比重	农产品电商	羊肉产量	羊肉消费
内蒙古	156	20.39	21	1	2
上海	96	12.55	4	31	1
江苏	94	12.29	1	16	3
山东	63	8.24	6	3	4
甘肃	57	7.45	27	7	15
广东	38	4.97	3	30	5
宁夏	33	4.31	9	28	9
北京	29	3.79	24	13	11
新疆	26	3.40	2	25	27
浙江	25	3.27	26	2	12
安徽	19	2.48	13	5	16
四川	17	2.22	14	18	19
河北	16	2.09	18	4	21
陕西	15	1.96	5	8	17
河南	11	1.44	10	6	8
青海	11	1.44	15	15	7
山西	10	1.31	30	11	23
湖北	9	1.18	23	19	18
辽宁	8	1.05	8	14	13
福建	7	0.92	29	21	26
黑龙江	5	0.65	22	10	30
重庆	4	0.52	25	22	28
湖南	4	0.52	7	24	22

广西	3	0.39	12	12	24
贵州	2	0.26	17	27	20
江西	2	0.26	19	20	14
海南	2	0.26	16	23	29
云南	1	0.13	11	9	10
天津	1	0.13	28	29	31
西藏	1	0.13	31	17	6
吉林	0	0.00	20	26	25
合计	765	100			

资料来源:电商排名是根据阿里研究院公布的《从“客厅革命”到“厨房革命”——阿里农产品电子商务白皮书 2016》中 2016 年的农产品电商销售额进行的排序;产量排名是根据《2017 中国统计年鉴》中 2016 年各省份的羊肉产量进行的排序;消费量排名是根据《2017 中国统计年鉴》中 2016 年各省份的消费量进行的排序;单位(个、%)。

表 6 淘宝羊肉产品销售总量和总额

销量范围	产品数量	所占比重	销量均值	总销量
0—10	1349	67.62%	3.55	4784.00
10—30	362	18.15%	17.82	6450.00
30—50	92	4.61%	40.03	3683.00
50—100	108	5.41%	71.27	7697.00
100—200	45	2.26%	137.20	6174.00
200—500	31	1.55%	318.94	9887.00
500—800	4	0.20%	596.50	2386.00
800 以上	4	0.20%	1259.25	5037.00
总体	1995	100.00%	23.11	46098.01

注:总销量=销量均值*产品数量,单位(个、单)

五、实证方案及结果

根据 SVM 方案的设定,按照训练集和测试集样本量的差异,包括训练集为 3/4 样本的 SVM 分类结果和训练集为 4/5 样本的 SVM 分类结果;每个表中根据涉及的变量情况,分为方案 1-6;同时调整成本约束参数 C,使其分别为 $c=1$ (R 软件中 SVM 的默认值), $c=100$ (SVM 官方说明书应用最广泛的参数值),以及 $c=10000$ (按照默认值和广泛参数值同等倍数放大的参数值),最终结果如表 7 和表 8 所示,基于两个表格显示的分类结果,对数据进行横向以及纵向对比。

从约束参数的角度,当 $c=1$ 时,由于允许误差很大,所以在剔除不同的变量以后,对于分类结果的影响并不是很大;而随着 c 增加至 100 和 10000 时,分类的精确度提高,且 $c=100$ 和 $c=10000$ 的结果差别不大,故在此将 $c=100$ 的结果作为主要分析对象。无论是 3/4 样本还是 4/5 样本的六个方案结果都反映出以下四点共同的结论,从而判断出羊肉电商销售的影响因素及其重要程度。

1. 当考虑所有特征指标时, SVM 模型的分类精确度是最高的, 即当删除其他任何类别的特征指标, 分类的精确度均会下降, 这表明评论特征、交易特征、店铺特征、服务特征以及产品特征都会影响到产品的销量;

2. 当删除评论特征指标时, SVM 模型的分类精确度是最低的, 即删除评论特征以后的分类精确度低于删除其他四类特征指标的精确度, 这反映出相比于其他四类变量, 评论特征对于产品销量的影响作用更为显著;

表 7 训练集为 3/4 样本的 SVM 分类结果

方案	变量情况	c=1	c=100	c=10000
方案 1	全部变量	75. 75%	75. 15%	75. 15%
方案 2	删除评论特征	74. 15%	66. 53%	67. 74%
方案 3	删除交易特征	75. 15%	73. 55%	73. 55%
方案 4	删除店铺特征	75. 95%	74. 55%	73. 75%
方案 5	删除服务特征	75. 15%	73. 55%	73. 95%
方案 6	删除产品特征	75. 95%	74. 35%	74. 15%

3. 当删除店铺特征指标时, SVM 模型的分类精确度要高于删除其他四类特征指标的精确度, 这一定程度上表示, 其他变量对淘宝平台成交量的影响比店铺特征更为显著;

4. 通过第二和第三点可以看出, 交易特征、服务特征以及产品特征对产品销量的影响程度介于评论特征和店铺特征之间。

5. 除了已涉及的这五类指标, 势必存在另外的影响因素也影响到销量。虽然五类指标已足以预测销量形成的 75%, 但仍有 25%的空间需要在未来研究中不断探索。

表 8 训练集为 4/5 样本的 SVM 分类结果

方案	变量情况	c=1	c=100	c=10000
方案 1	全部变量	75. 50%	75. 25%	75. 25%
方案 2	删除评论特征	74. 50%	65. 75%	67. 00%
方案 3	删除交易特征	74. 75%	72. 75%	72. 75%
方案 4	删除店铺特征	75. 00%	73. 50%	72. 75%
方案 5	删除服务特征	75. 25%	73. 00%	73. 50%
方案 6	删除产品特征	75. 75%	71. 75%	71. 50%

SVM 的结果分析表明, 在线评论、交易过程中是否提供相关的保障承诺、服务的水平高低、产品价格以及包装是否满足消费者偏好、店铺的运营资质等方面均是影响在线消费决策的重要因素, 而其中在线评论的影响力更为显著。这反映出不完全信息条件下, 消费者在线购买生鲜产品时, 会更倾向于参考在线评论来进行消费决策。该结论与王英等^[20]的研究具有一致性, 都反映出在线评论显著影响生鲜电商的农产品销量。传统意义上的农产品购买行为都发生在农贸市场或超市, 由消费者通过看、摸、闻、尝等方式形成自我感知, 进而结合商品定价, 判断是否值得购买。而发展羊肉等生鲜电商, 在线购买农产品, 现实环境的缺失促使消费者更多地依赖于平台上其他消费者购买后的体验, 以及其反映到在线评论中的图片和文字说明, 从而做出与传统交易方式类似的判断结果。

当然，坏单包赔以及质量保障等承诺也会降低消费者所承担的风险，从而提高其对产品的心理预期。王克喜和戴安娜^[9]的研究曾指出，网购平台信息、系统、送货和服务质量对网购生鲜农产品意愿具有显著正向影响，该结论与本研究结果较为类似。互联网时代下，消费者更加看重消费的服务体验以及产品的可选择性，换句话说，电商平台向消费者提供的服务体验以及产品的丰富度，在一定程度上提高了消费者的效用水平，因而无论是服务水平的高低还是包装大小，都会影响到消费者的购买决策。此外，电商卖家是否值得信任，也是保证产品质量的重要方面，消费者在购买过程中也势必将其作为参考的依据之一。

六、政策建议及研究展望

理论上，平台型生鲜电商的农产品销售行为对于“小农户对接大市场”，提高农户福利具有重要的现实意义，而现实中农户、政府和企业也正积极探索生鲜电商的经营模式，尝试多种手段来提高生鲜电商发挥的实际作用。但笔者的研究显示，当前以羊肉为代表的生鲜电商规模仍然有限，基于研究结果，不同主体均可以采取更为细致的措施来促进未来生鲜电商的发展。

首先，对在平台上销售生鲜产品的农户或个体户而言，需要通过改善其经营策略增进销售。一方面，充分利用平台反馈的数据，以销定产，制定更加合理的包装规格和单品价格等，从而更好地满足消费者个性化、多样性的产品需求。另一方面，可以考虑提供适当合理的保障措施和交易福利，例如坏单包赔、质量保障、较低的快递费用等，降低消费者在交易过程中承担的风险，提高其对产品的心理预期。此外，虽然经营的商品不是大品牌，但仍应注意品牌的打造，实现差异化的销售策略。

其次，政府作为政策的制定者，应当继续鼓励支持生鲜电商的发展，以改善现阶段生鲜电商规模仍处于较低水平的现状。当然，值得注意的是，应根据不同的省份优势，进行政策引导与支持，对电商环境发展良好的、消费基础稳定的、农业产业优势强大的省份和地区的政策和扶持侧重点都应有所不同。对于产品质量本身，可以通过制定生鲜产品分类标准，即给出有关其大小、颜色、味道、营养成分等性状的具体差异，针对差异设定标准细则的分类办法，来降低消费者购买风险，从而促进消费决策。

最后，第三方平台作为沟通供需双方的重要中介，应当采取适当的措施，促进信息的交流，减少信息不对称，才能更好地达成交易。在农产品质量安全日益成为社会公众的核心诉求时，电商作为一种销售渠道在销售产品时，能够通过平台提供的互动机制，为消费者提供更多的农产品质量信息。为了更好地联结买家和卖家，平台一方面可以进一步优化在线评论界面，以方便消费者的阅读，提供更多的信息、增进消费者对产品信息的掌握程度；另一方面，可以提高规范化店铺入驻管理，例如担保资金等，也可适当扩大物流水平等评分差距，为消费者提供更多的质量信号。

笔者将大数据方法具体应用到电子商务的经济学研究中，为未来的研究提供了可借鉴的研究方法，以及下一步研究的思路。研究方法上，无论是 Python 爬虫的挖掘方法还是 SVM 的分析方法，都属于大数据分析方法的范畴，利用大数据方法，既可以拓展经济研究的数据来源，对公开数据以及调查问卷等传统数据获取方法进行补充；又能够在数据处理过程中引入与传统的统计回归方法不同的算法，例如人工智能、支持向量机、机器学习算法、神经网络等，从而更有效地从海量半结构化、非结构化数据中探索经济关系。该研究方法可以推广至其他电商平台以及农产品的数据采集和销量影响因素的分析上。本文研究表明不同类别的因素对于销量的影响程度不尽相同，每一类别的因素又包含更为细致的影响指标，下一步研究将进一步深化生鲜电商交易因素的分析，可以重点探析在线评论及其具体特征对农产品在电商平台上交易的作用效果。

参考文献:

- [1]李华杰，史丹，马丽梅．基于大数据方法的经济研究:前沿进展与研究综述[J]．经济学家，2018(6):96-104.
- [2]夏晓平，李秉龙．品牌信任对消费者食品消费行为的影响分析——以羊肉产品为例[J]．中国农村观察，2011(4):14-26+96.

-
- [3]Canavari M.,Fritz M.,Hofstede G.J.,Matopoulos A.,Vlachopoulou M.The role of trust in the transition from traditional to electronic B2B relationships in agri-food chains[J].Computers&Electronics in Agriculture,2010(2):321-327.
- [4]侯振兴. 区域农户农企采纳农产品电子商务的影响因素[J]. 西北农林科技大学学报(社会科学版), 2018(1):66-74.
- [5]林家宝, 李婷, 鲁耀斌. 环境不确定性下农产品电子商务能力对企业绩效影响的实证研究[J]. 商业经济与管理, 2018(9):64-75.
- [6]葛继红, 王文昊, 汤颖梅. 我国生鲜电商发展模式及其适用条件分析[J]. 贵州社会科学, 2018(1):133-138.
- [7]曾亿武, 翟李琴, 郭红东. 电子商务视阈下的农户诚信经营意识:强化还是弱化?——来自江苏沭阳的证据[J]. 商业经济与管理, 2019(5):15-25+45.
- [8]Ernst S.,Hooker N.H.Signaling Quality in an E-Commerce Environment[J].Journal of Food Products Marketing, 2006(4):11-25.
- [9]王克喜, 戴安娜. 基于 Logit 模型的绿色生鲜农产品网购意愿的影响因素分析[J]. 湖南科技大学学报(社会科学版), 2017(2):87-93.
- [10]梁文卓, 侯云先, 葛冉. 我国网购农产品特征分析[J]. 农业经济问题, 2012(4):40-43.
- [11]吴自强. 生鲜农产品网购意愿影响因素的实证分析[J]. 统计与决策, 2015(20):100-103.
- [12]Cavallo A,Rigobon R.The Billion Prices Project:Using Online Prices for Measurement and Research[J]. Journal of Economic Perspectives,2016(2):151-178.
- [13]雷怀英, 王毓彬, 吴英明. 基于网络检索数据的 CPI 预测研究[J]. 经济问题, 2018(10):22-26.
- [14]赖胜强, 朱敏. 网络口碑研究述评[J]. 财贸经济, 2009(6):127-131.
- [15]卢向华, 冯越. 网络口碑的价值——基于在线餐馆点评的实证研究[J]. 管理世界, 2009(7):126-132+171.
- [16]殷国鹏. 消费者认为怎样的在线评论更有用?——社会性因素的影响效应[J]. 管理世界, 2012(12):115-124.
- [17]龚诗阳, 刘霞, 刘洋, 赵平. 网络口碑决定产品命运吗——对线上图书评论的实证分析[J]. 南开管理评论, 2012(4):118-128.
- [18]汪旭晖, 聂可昱, 陈荣. “解释行为”还是“解释反应”?怎样的在线评论更有用——基于解释类型的在线评论对消费者购买决策的影响及边界条件[J]. 南开管理评论, 2017(4):27-37.
- [19]陶晓波, 张欣瑞, 等. 在线评论、感知有用性与新产品扩散的关系研究[J]. 中国软科学, 2017(7):162-171.

-
- [20]王英, 杨丹, 王乐, 张卓, 谢鹤鹏, 郝会成. 基于品牌效应的在线评论对稻米销量影响的实证研究[J]. 中国农业大学学报, 2018(1):178-187.
- [21](德)雅姆尔, (奥)理查德·劳森. 用Python写网络爬虫[M]. 李斌, 译. 北京:人民邮电出版社, 2018:11-12.
- [22](美)米切尔. Python网络数据采集[M]. 陶俊杰, 陈小莉, 译. 北京:人民邮电出版社, 2016:2-10.
- [23]孙易冰, 赵子东, 刘洪波. 一种基于网络爬虫技术的价格指数计算模型[J]. 统计研究, 2014(10):74-80.
- [24]阮素梅, 何浩然, 李敬明. P2P 借贷中借款人的违约风险评估——基于“人人贷”数据的实证分析[J]. 经济问题, 2017(12):45-50+99.
- [25]Ye Q., Law R., Gu B. The impact of online user reviews on hotel room sales[J]. International Journal of Hospitality Management, 2009(1):180-182.
- [26]Ma Y., Xiang Z., Du Q., et al. International Journal of Hospitality Management Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep learning[J]. International Journal of Hospitality Management, 2018(71):120-131.
- [27]霍红, 张晨鑫. 在线评论质量丰富度对购买异质性产品的影响[J]. 企业经济, 2018(6):77-83.
- [28]Cortes, C., Vapnik, V. Support-vector network[J]. Machine Learning, 1995(20):1-25.
- [29]薛宁静. 多类支持向量机分类器对比研究[J]. 计算机工程与设计, 2011(5):1792-1795.
- [30]Cernadas E., Amorim D. Do we need hundreds of classifiers to solve real world classification problems?[J]. Journal of Machine Learning Research, 2014(1):3133-3181.
- [31](美)兰茨. 机器学习与R语言[M]. 李洪成, 许金炜, 李舰, 译. 北京:机械工业出版社, 2016:161-166.
- [32]Meyer D. Support Vector Machines: The Interface to libsvm in package e1071[EB/OL]. (2019-03-18) [2019-03-18]. <https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf>
- [33]Yeh, I. C. Modeling of strength of high-performance concrete using artificial neural networks[J]. Cement & Concrete Research, 1998(12):1797-1808.
- [34]Frey P. W., Slate D. J. Letter recognition using Holland-style adaptive classifiers[J]. Machine Learning, 1991(2):161-182.