E-Commerce Letters 电子商务评论, 2025, 14(11), 308-315
Published Online November 2025 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2025.14113437

Hans汉斯

基于机器学习的休闲零食电商营销策略优化

徐媛彬，彭定涛*

贵州大学数学与统计学院，贵州 贵阳

收稿日期：2025年9月19日；录用日期：2025年10月10日；发布日期：2025年11月10日

摘　要

随着电商平台成为休闲零食行业的核心销售渠道，挖掘消费者需求与优化营销策略成了解决行业供需衔接问题的关键部分。本文以休闲零食电商消费数据作为研究对象，通过整合多源数据创建研究体系：首先采用网络爬虫获取电商平台消费者的评论数据，再加上线上问卷数据补充消费偏好信息；其次通过与自编码器结合的随机森林算法识别消费者的首要关注因素，采用逻辑回归模型来量化变量对购买频率的影响；最后按照K-means聚类将消费者划分为四类群体，给出策略。研究结果可为休闲零食企业制定电商营销策略、提升市场竞争力提供数据支撑。

关键词

消费者行为，数据挖掘，营销策略优化

Optimization of E-Commerce Marketing Strategy for Leisure Snacks Based on Machine Learning

Yuanbin Xu, Dingtao Peng*

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: September 19, 2025; accepted: October 10, 2025; published: November 10, 2025

Abstract

With e-commerce platforms becoming the core sales channel for the leisure snacks industry, uncovering consumer demand and optimizing marketing strategies have become key to resolving supply-

*通讯作者。

文章引用: 徐媛彬, 彭定涛. 基于机器学习的休闲零食电商营销策略优化[J]. 电子商务评论, 2025, 14(11): 308-315.
DOI: 10.12677/ecl.2025.14113437

demand mismatches. This study takes e-commerce consumption data of leisure snacks as its research object and constructs a comprehensive research framework by integrating multiple data sources. Firstly, web crawler technology is used to obtain consumer review data from e-commerce platforms, and online questionnaire data is combined to supplement information on consumer preferences. Secondly, the random forest algorithm combined with an autoencoder is applied to identify the primary concerns of consumers, and the logistic regression model is used to quantify the impact of variables on purchase frequency. Finally, consumers are divided into four groups based on K-means clustering, and targeted strategies are proposed. The research results can provide data support for leisure snack enterprises to develop e-commerce marketing strategies and enhance market competitiveness.

Keywords

Consumer Behavior, Data Mining, Marketing Strategy Optimization

1. 引言

在数字经济迅速发展的大环境之下，电商平台依靠其便捷性以及具有丰富多样的商品种类，已然成为休闲零食行业销售的关键渠道。从市场实际情况来看，消费者借助电商平台来挑选薯片、坚果、饼干这类休闲零食的行为变得越来越常见，这种趋势促使行业规模不断扩大[1]。

然而当下行业发展状况与消费需求之间依然存在着较为突出的矛盾[2]：一方面，多数休闲零食企业在开展电商运营时，不是依靠同质化产品进行布局，就是欠缺对消费者偏好的深入了解，最终使得营销资源投入和实际转化效果出现了脱节的情况；另一方面，电商场景的“无法线下体验”，致使消费者难以直观感受到零食的口味、口感等属性，只能借助图片、文字评论来间接做出判断，容易因为信息不对称而产生购买决策方面的顾虑，对购买意愿以及复购频率造成影响。

依据现有的相关研究情况而言，对于休闲零食消费所展开的探讨，大多聚焦于产业供应链的优化方面，或者是针对线下消费场景进行分析[3]，然而针对电商场景所开展的专项研究，明显存在着不足之处：有一部分研究仅仅是借助定性描述来归纳消费特征，缺少量化数据作为支撑，没有构建起“数据收集－需求识别－行为分析－策略输出”这样完整的逻辑链条，以至于很难直接为企业的电商营销提供有实际可操作性的指导。

鉴于此，本文围绕休闲零食电商消费数据展开研究，构建了多源数据整合的研究体系，借助网络爬虫技术收集电商平台用户评论数据，并结合线上问卷补充消费偏好信息，以此为研究搭建数据基础。再运用随机森林、逻辑回归、K-means 聚类等机器学习方法，识别消费者核心关注因素以及量化变量对购买频率的影响。依据不同群体特征提出营销策略优化建议，此研究成果可为休闲零食企业定位目标客群、提升电商竞争力提供数据支持。

2. 数据收集

在电商经济渗透休闲零食行业之际，消费者在电商平台的评价，问卷反馈等数据，是了解其需求与偏好的关键资源。本章节主要围绕电商场景下的休闲零食数据搜集展开，为后续分析提供数据基础。

2.1. 电商平台文本挖掘

使用网络爬虫技术，爬取电商平台中休闲零食的评论信息，涵盖不同定位、不同规模的零食品牌。评论信息主要包括评论内容、用户、评论时间等，对评论内容进行分析，共收集到休闲零食商品相关评论 10,000 条，最终爬取的评论信息部分如表 1 所示。

Table 1. Partial comment information
表 1. 部分评论信息

用户	评论内容	评论时间
a*k	买了好多次了，小包装拿着方便，喜欢吃这种，包装很精致，说明下了功夫的，饼干每片酥脆酥脆的奶味特别浓郁，不甜关键味道也好！	2025.2.15
真*儿	特别好吃一下子吃了五六袋，味道挺好，单独包装，真是物超所值。	2025.8.16

爬取的评论数据，可能存在虚假用户评论、评论与该主题毫无关联等现象，所以对原始数据进行预处理。首先处理基础数据问题：对于购买频次这类数值型数据，若缺失比例低于 5%，采用均值插补；缺失比例在 5%~20%之间，采用 K 近邻算法插补；若缺失比例超过 20%，且非核心变量，直接剔除。对于情感倾向标签这类分类型数据，采用众数插补，同时借助哈希算法生成唯一标识，通过比对删除重复记录，避免数据冗余。

针对虚假用户评论这类噪音数据，用滑动窗口算法统计 1 小时内用户的评论次数，若某用户发布 5 条及以上评论，且评论内容相似度超过 0.8，就标记为虚假用户评论。最后剔除虚假用户评论及关联数据，保证留下的评论是真实有效的，为后续分析提供数据支撑。

分词是文本分析的重要部分，其原理为将一句话按照规则划分为单个中文词语，再统计词语出现的频次、频率等信息。本文中的分词工作使用 python 语言中的 jieba 库完成[4]，对评论数据进行文本分词，使用高频词绘制词云图来可视化文本，词云图如图 1 所示。

Figure 1. Word cloud of comment data
图 1. 评论数据词云图

词云图中，图中字体的大小与词频息息相关，字体越大代表词频越高。可以看出，影响消费者购买零食的高频词语主要是包装、质量、物流等。可以认为，以上几个因素是消费者是否购买休闲零食的主要原因。

2.2. 问卷调查分析

除了对文本评论进行网络爬虫采集外，还采用线上形式收集与休闲零食相关的调查问卷，这种方式可以从更为广泛的角度获得消费者对休闲零食的看法。再与文本挖掘得到的内容相结合，获得消费者对休闲零食的偏好，并更加全面地了解休闲零食在电子商务上的发展状况，为后续研究的展开提供数据信息。

问卷的调查对象为休闲零食的线上消费者，在设计调查问卷时遵循问卷设计的目的性、明确性等相关原则。设计的调查问卷主要有三个部分：基本信息，休闲零食的市场容量，消费者购买认知。调查一共收集到问卷 1163 份，其中有效问卷 1002 份。对问卷中的量表进行信度检验，并利用 SPSS 软件计算 Cronbach’s Alpha 系数作为评价标准，得到信度系数为 0.883，表明量表通过信度检验。再通过 KMO 和 Bartlett 球形检验因子，进行结构效度分析，得到 Bartlett 检验对应的 P 值小于 0.05、KMO 值为 0.893，故问卷结果具有可靠性。

根据问卷与消费现状，线上购买过休闲零食的人群占比高，反映其可观的市场潜力。问卷显示，女性受访者占 61.18%，男性占 38.82%，女性占比更高；休闲零食消费者中，每月线上花费多集中于 21~100 元。在购买原因方面，39%因“方便快捷”购买，21%因“价格实惠”购买，19%因“选择多样”而购买等，可为品牌营销提供方向。

3. 消费者关注度及行为分析

电商平台的众多数据为剖析消费者对休闲零食的关注度与偏好提供了可能。通过随机森林和逻辑回归算法，从电商数据中找到关键的影响要素，就能够准确掌握电商环境之下消费者做出决策的思路，从而优化企业的电商营销手段。

3.1. 自编码器与随机森林相结合的消费者关注度分析

要分析消费者的关注因素，实际上就是通过随机森林分析特征与消费者之间的关系。文章[5]和[6]将自编码器与随机森林结合，避免了传统特征选择的单一性，使随机森林拥有优异的性能。

自编码器作为无监督的神经网络，在保持重构误差尽可能小的情况下进行特征提取。从结构上看，自编码由输入层、隐藏层和输出层构成，隐藏层的层数往往小于输入层，这可以实现对输入数据的降维，解决输入数据特征冗余的问题，适合对高维，有冗余的电商数据进行处理。在分析消费者关注度时，关于消费者对休闲零食关注的行为、偏好等数据，往往维度高且存在大量冗余信息，自编码器能有效过滤这些冗余信息，提取出更具代表性的核心特征，为后续分析奠定基础。

随机森林通过学习给定的经验数据，从大量变量中找出重要的变量，它是一种基于决策树的分类器集成算法，可以解释二分类变量 Y 受到自变量的影响程度[7]。在对消费者关注度的分析里，若只靠随机森林来选特征，很可能会因为单棵决策树出现过拟合的情况，或者特征选择角度单一等问题，使得选出来的特征不能全面又精准地体现消费者关注因素。但把自编码器和随机森林结合后，自编码器先对数据做特征提取和降维，为随机森林提供更精简、更核心的特征集合。这样使得随机森林在分析消费者关注度时，从多个决策树的不同角度去综合考虑，避免了传统单一特征选择方法的片面性，让最终得到的消费者关注因素分析结果更全面、准确。因此，先采用自编码器处理数据。将输入层和输出层设置为 1002，隐藏层设置为 600，对 1002 位受访者的数据进行训练，提取隐藏层的数据用于随机森林进行关注度分析。

通过随机森林模型，对影响消费者购买休闲零食产品的品牌口碑或形象、口味是否达到预期、食品安全、原料品质、包装良好这 5 个因素的重要程度进行分析与排序，得到如表 2 所示的结果。

Table 2. Attention ranking chart
表 2. 关注度排名图

影响指标	排序
口味是否达到预期	1
品牌口碑或形象	2
原料品质	3
食品安全	4
包装良好	5

由表 2 可得，在电商场景下，消费者对休闲零食的关注因素里，关注度最大的是口味是否达到预期，这在电商经济的背景下有以下成因：

(1) 产品价值属性起着决定性作用：休闲零食的核心消费需求在于满足味觉愉悦，并非饱腹以及营养等需求。与包装的外观辅助作用相比，口味直接决定了消费体验的核心质量，要是口味未达到预期，即便其他因素存在优势，也难以弥补消费者满意度的缺失，消费者会优先将注意力集中在这一有决定性的指标上。

(2) 电商场景会使口味信息不对称的问题放大：在线下进行购物的时候，消费者可借助试吃，以及观察零食酥脆程度、软糯程度等方法，直接对口味做出判断，然而在电商场景当中，消费者仅仅只能依靠图片、文字描述以及他人评价来间接感知口味，存在较为十分突出的“体验滞后性”，这样的信息差会让消费者对于“口味是否符合自身偏好”产生更为强烈的不确定性，在做决策的时候会优先去关注口味是否可达到预期，以此来降低买错以及浪费的风险。

3.2. 基于逻辑回归的用户特征及行为分析

逻辑回归帮助了解不同因素对消费者购买行为产生的影响程度，可以更好地去预测和解释消费者行为，制定针对性的市场策略[8]。在电商经济蓬勃发展的大背景下，对消费者购买频率的影响因素进行分析是有必要的。

聚焦消费者购买休闲零食频率的影响因素，建立二元逻辑回归模型。将月平均购买频率设置为响应变量，将月平均购买频率 2 次及以下定义为购买频率低，购买频率 3 次及以上定义为购买频率高，选取性别(Gender)、职业(Occupation)、月收入(Income)、方便快捷(Quick_easy)、多样化的选择(Diverse options)、包装(Wrap)、价格实惠(Affordable)、品牌(Brand)、口味(Taste)和食品安全(Safety)，这 10 个指标作为模型的因变量。使用极大似然估计方法初步建立的模型，对模型进行分析。对不显著的变量，考虑 AIC 准则和逐步回归法对自变量进行筛选，进一步优化模型，最终得到的模型结果如表 3 所示。

Table 3. Coefficient table of logistic regression model
表 3. 逻辑回归模型系数表

变量名称	回归系数	标准差	P 值	显著性
Intercept	0.04360	0.29964	0.884299	
Income	0.14954	0.04414	0.000705	***
Diverse options	0.65269	0.14996	1.35E−05	***
Wrap	0.54139	0.16604	0.001112	**
Affordable	−0.30529	0.17987	0.089636	.
Brand	0.11806	0.08312	0.125493	
Safety	0.17471	0.06221	0.004976	**

注：用于标注逻辑回归模型各变量系数的统计显著性水平，***表示 P 值 < 0.001，**表示 P 值 < 0.01，.表示 P 值 < 0.1，无星号表示 P 值 ≥ 0.1。

在电商场景下消费决策受特殊因素影响，传统严格标准易遗漏关键变量。结合随机森林得到的结果，品牌口碑或形象是消费者关注的第二大因素，但是逻辑回归的 P 值却大于 0.05，为了能更全面地反应消费者对休闲零食的消费偏好，将逻辑回归模型的显著性水平从传统的 0.05 放宽至 0.1。对比全模型和 AIC 准则的样本 AUC 值，全模型为 0.6686，AIC 模型为 0.6661，差别不大。使用 AIC 准则选择后的模型变量只有一个不显著，ROC 曲线也与全模型类似，仍保持良好预测稳定性。本文最终建立的模型为：

$$\ln\frac{p}{1-p} = 0.0436 + 0.1495\text{Income} + 0.6527\text{Diverse options} + 0.5414\text{Warp} - 0.3053\text{Affordable} + 0.1181\text{Brand} + 0.1747\text{Safety}. \tag{1}$$

从模型结果公式(1)可知，在电商场景下，月收入、多样化的选择、产品包装、价格优惠、品牌、食物安全这几个变量在显著性水平为 0.1 的情况下均显著，即上述变量都会对消费者购买休闲零食的频率造成显著影响。结合电商经济的特点，可得出以下结论：

(1) 在电商平台的消费环境中，消费者的性别和职业对休闲零食的购买频率影响甚微。电商平台打破了传统消费场景下性别和职业带来的消费限制，为不同性别、职业的消费者提供了相对平等、便利的休闲零食购买渠道。

(2) 休闲零食在电商平台上提供多样性的选择、良好的包装、实惠的价格对消费者购买频率的影响较大。电商平台有着丰富的展示空间和便捷的搜索功能，让消费者能更轻松地接触到多样化的休闲零食产品；而精美的包装不仅能在物流运输中更好地保护产品，还能通过电商平台的图片展示吸引消费者；价格实惠更是电商平台促销活动的卖点，让消费者产生购买的欲望。

基于上述结论，提供多样化的选择、设计符合消费者审美的包装、优惠的价格以及提高食品安全性都会提高消费者购买休闲零食的频率。电商平台的用户评价、直播带货等功能，可以令消费者更加直观地了解到休闲零食的多样性和包装情况；价格优惠信息通过电商平台的推送能够快速传达给消费者；对于食品安全相关的检测报告、认证标识等可以在电商平台上进行展示，增强消费者对休闲零食的信任，从而促进购买。

4. 基于 K-Means 客户群体分类的营销策略

电商经济的发展让休闲零食消费行为呈现多样性。针对电商平台消费者购买频率、聚类分析，可清晰划分市场群体，这对休闲零食企业在电商场景下实现精准营销、提升市场竞争力意义重大。

4.1. 线上消费者群体聚类分析

在电商经济蓬勃发展的当下，休闲零食市场发展势头正盛。为了探究不同类型的线上消费者群体对于休闲零食的态度与看法，对问卷数据进行分析。通过 K-means 聚类按照基本特征分类，将相同类别的个体尽可能地归为一类，使不同类别的个体间存在较大的差异，便于根据不同的类别的特征进行分析和识别，减少数据分析的维度[9]。

对受访者进行 K-means 聚类前，进行变量选择。选取表 4 的两个变量作为聚类依据。

Table 4. Clustering variable summary table
表 4. 聚类变量汇总表

变量	涉及的相关问题	变量定义
频率	购买休闲零食的频率是平均每月几次	1分：2次及以下 2分：3 ~ 6次 3分：7 ~ 9次 4分：10次以上

续表

消费	平均每个月花多少钱购买休闲零食	1分：20元及以下 2分：20～50元 3分：51～100元 4分：101元及以上

对受访者的购买频率和总消费，使用 R 语言进行聚类分析，将客户分为四类。并参考 RFM 对于分类客户的名称和当前市场常见的用户分类名称，对这四类客户命名为活跃客户、发展客户、潜在客户和沉默客户[10]。最终得到结果如表 5 所示。

Table 5. Group classification characteristics
表 5. 群体分类特征

名称	活跃客户	发展客户	潜在客户	沉默客户
购买频率	高(4 分)	一般(2 分)	一般(2 分)	低(1 分)
总消费	高(4 分)	高(3 分)	一般(2 分)	低(1 分)
受访者个数	132	284	216	253

通过统计不同客户类型线上购买休闲零食的原因，得到结果如图 2 所示。

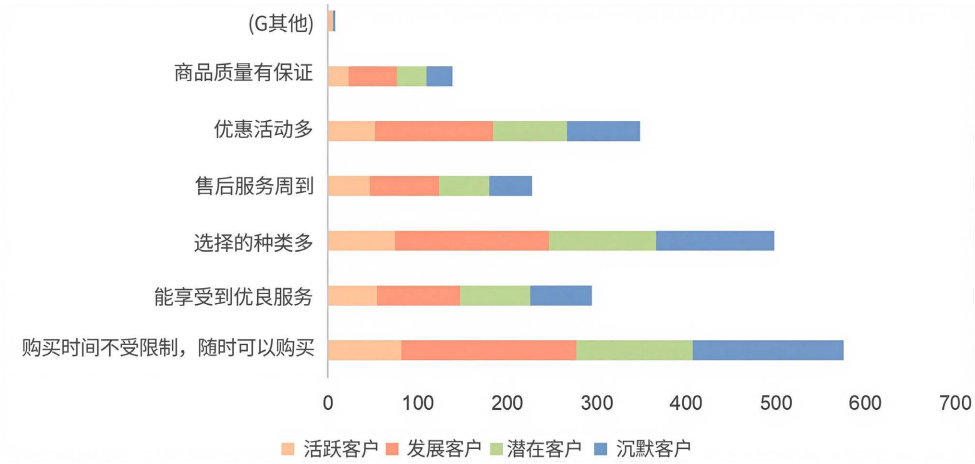

Figure 2. Reasons for different types of customers to purchase leisure snacks online
图 2. 不同客户类型线上购买休闲零食的原因

由图 2 可知，购买时间的灵活性和休闲零食种类多是四类客户选择线上购买的主要原因。活跃客户对优良服务和售后服务也有较高的期待，可通过优质的售后服务来巩固其忠诚度。发展客户对优惠活动表现出极大的兴趣。潜在客户和沉默客户虽然在购买时间的灵活性上显示出较高的兴趣，但在其他方面的关注度较低。同时，四类客户对价格的敏感度不同。发展客户和沉默客户关注产品的质量，单纯的低价策略可能无法吸引，需要在保证质量的同时，给出合理的价格。特别是沉默客户及其重视价格因素的影响。

4.2. 基于线上消费场景的客户体验优化策略

通过深入分析不同客户群体的特征和偏好，针对休闲零食提出可行的营销策略：

(1) 对于活跃客户，以优良服务巩固忠诚度。强化售后服务对于提高客户忠诚度和复购率至关重要。创建快速响应机制，及时解决客户的咨询和投诉，保证客户的满意度，定期跟进客户反馈，通过顾客满意度调查等手段，了解客户的需求，持续改善产品和服务质量。同时，优化用户界面，简化购物流程。

(2) 对于发展客户，以优惠价格促进消费。推出多样化的组合套餐，采取灵活地定价，例如，节假日促销，会员折扣活动等。并且开展积分奖励计划，以“邀请好友下单”、“购买积分兑换礼品”等方式，促进消费，提升购买频率。

(3) 对于潜在客户，以“体验”降低门槛。采用“包邮退货退款”和“免费试吃”的方案，在客户收到产品后，可以免费试吃 1~2 份，若不满意可以退还剩余的商品并全额退款，降低“尝试成本”，引导客户消费。采用“首单减 5 元”等优惠，降低首次消费门槛，吸引具有观望心理的消费者。

(4) 对于沉默用户，降低决策门槛，激活消费。推出“小分量”的组合装零食，设置较低的单价，降低单次购买成本。根据浏览记录，精准推送商品。通过手机短信发放专属消费卷和复购卷，例如“本月第二次消费 8.5 折”等活动，激励消费，培养消费习惯。

电商经济给休闲零食行业的发展提供了广阔的空间，精准掌握不同类型客户的需求，制定出针对性的营销策略，可以有效提高休闲零食在电商平台上的销量和市场份额，推动休闲零食行业的不断发展。

5. 总结

本文以休闲零食电商消费数据为核心，通过网络爬虫获取电商评论，再配合线上问卷创建多源数据体系，利用随机森林，逻辑回归，K-means 聚类等方法展开研究，发现食品安全是消费者最重视的因素，并且量化了月收入，多样化选择等对购买频率的影响，把消费者分成四类群体，最后给出差异化策略，给休闲零食企业制定电商营销策略给予支撑，也丰富了食品电商领域消费者行为研究的范式。

基金项目

国家自然科学基金项目(12261020)、贵州省科技计划项目(黔科合基础-ZK[2021]009)和贵州省高层次留学人才创新创业择优资助重点项目([2018]03)。

参考文献

[1] 江一苇, 周琦. 三只松鼠回暖[J]. 21 世纪商业评论, 2024(11): 22-24.

[2] 黄译莹. 休闲零食企业盈利模式研究——以良品铺子为例[J]. 商场现代化, 2025(9): 20-22.

[3] 王卓. 供应链数字化转型对休闲食品企业绩效影响的研究[D]: [硕士学位论文]. 呼和浩特: 内蒙古财经大学, 2025.

[4] 黄强. 基于文本挖掘的汽车用户需求分析[J]. 汽车与新动力, 2024, 7(1): 80-84.

[5] Lin, T. and Jiang, J. (2021) Credit Card Fraud Detection with Autoencoder and Probabilistic Random Forest. *Mathematics*, **9**, Article No. 2683. https://doi.org/10.3390/math9212683

[6] 闫蒙蒙, 陈建凯, 孟会贤, 王鑫. 随机森林与自编码器相结合的自适应特征选择算法[J]. 人工智能科学与工程, 2023(9): 39-47.

[7] 石佳鑫, 张之政. 基于随机森林算法的电动车销售影响因素的研究[J]. 营销界, 2022(10): 41-43.

[8] 万欣, 黄翔, 王甫志. 基于逻辑回归的数据中心网络流量预测[J]. 计算机应用, 2023, 43(S2): 152-156.

[9] 胡新海, 叶建龙, 盛君贤. 基于 K-Means 聚类分析法的大数据环境下电商精确营销策略[J]. 廊坊师范学院学报(自然科学版), 2023, 23(4): 50-52.

[10] 吴花平, 冯薇薇, 李林. 基于 RFM 的聚类算法在零售市场客户细分研究[J]. 重庆理工大学学报(社会科学), 2024, 38(10): 138-149.