数据科学驱动的电商用户消费行为分析与应用

孙鹤诚

珠海科技学院，广东省珠海市，519041；

摘要：为了深入探究数据科学在电商用户消费行为分析领域里的应用价值，该研究运用数据科学方法，对用户行为数据的采集、预处理以及方法论运用等方面展开了细致入微的深入分析，成功构建了用户画像、消费行为预测模型以及个性化推荐系统，并将这些成果应用于营销策略的制定。通过对表格数据进行细致的关联分析，成功验证了所构建模型的科学性与合理性。研究结果表明，数据科学的应用能够切实有效地提升电商平台用户的体验感受，提高商业运作效率，进而增强其在市场中的竞争力。

关键词：数据科学；电商用户；消费行为分析；个性化推荐

DOI：10.64216/3080-1486.26.02.059

引言

数据治理与挖掘技术作为大数据时代的核心驱动力，已深度渗透至电商领域[1]。电商平台借助多个渠道，精心采集用户的各类行为数据，这些数据广泛涵盖了点击流、交易记录以及社交互动等不同的异构数据源，最终成功构建出一个高维且动态的用户行为数据库[2]。依托数据科学搭建起的分析框架，能够顺利完成用户分群工作，精准预测消费趋势，还能给出个性化服务推荐，进而为平台运营给予量化决策方面的有力支撑。用户行为分析模型构建需融合行为经济学理论与计算技术[3]。利用聚类分析的方法，细致地识别出用户的消费特征，再结合关联规则挖掘技术，深入揭示商品之间那些不易察觉的隐性关联，能够构建起精准且全面的用户画像体系。依靠时间序列分析搭建起的消费预测模型，可以实时动态地捕捉用户需求产生的变化，进而为库存的优化以及营销策略的调整提供可靠的科学依据。

1 数据科学在电商用户消费行为分析中的应用

1.1 数据科学的定义与电商领域的应用概述

数据科学是统计学、计算机科学以及应用数学这三者深度交融而形成的交叉学科，其关键之处在于借助系统性的数据分析方法，把数据里潜藏的模式与知识给揭示出来。在电商这个充满活力的领域里，数据科学的应用早已冲破了传统统计分析的条条框框，成功构建起一套以用户行为分析作为核心驱动力量的技术体系。该体系巧妙地整合了来自多个维度的数据资源，精心构建起用户行为预测模型，为电商平台的动态决策提供了坚实有力的支持。数据科学借助机器学习算法，对用户浏览路径、购买频次以及商品偏好等各类行为数据展开深度剖析，从而精准识别出用户的消费意图，在此基础上进一步优化商品推荐策略和库存管理流程。

1.2 用户行为数据的采集与预处理要点

在电商领域里，用户行为数据的采集搭建起了数据分析的基础架构，其数据来源包含了显性行为数据和隐性行为数据这两大主要类别。显性数据主要涵盖了用户自己主动产生的购买记录、对商品给出的评价、搜索时使用的关键词等这类结构化的信息；而隐性数据呢，则牵涉到用户在页面上的停留时长、鼠标移动的轨迹、把商品加入购物车的行为等非结构化的数据。数据采集技术告别了以往那种传统的日志记录方式，已逐步发展成实时流数据处理模式，借助分布式采集框架，能够达成对 PB 级数据的秒级快速响应。数据预处理在分析流程中占据着关键地位，其核心任务囊括了数据清洗、特征工程以及数据转换这几个重要方面。数据清洗主要是为了解决诸如缺失值填充、异常值检测以及重复数据剔除等数据质量问题，常用的方法有均值插补、KNN 邻近算法，还有基于规则的过滤机制等。特征工程借助特征选择、特征构造以及特征降维等手段，把原始数据转变为机器学习模型能够识别的特征向量形式。

1.3 数据科学方法论在用户行为分析中的具体运用

聚类分析，作为无监督学习领域里一种极具代表性的方法，在用户细分这一领域展现出了极为显著的价值。电商平台借助 K-means 算法或者 DBSCAN 密度聚类的

方法，能够把用户划分成不一样的群体，进而识别出高价值用户、价格敏感型用户以及潜在流失用户等细分市场。在消费趋势预测的领域里，时间序列分析扮演着至关重要的核心角色。当 ARIMA 模型和 LSTM 神经网络携手合作时，能够精准地预测出季节性商品在销售上的波动情况。某家服装电商巧妙地运用构建时间序列预测模型的方法，成功地将库存周转率提升了 19%，同时，还把缺货率降低了 12%。深度学习技术的巧妙引入，如同为分析工作打开了一扇新的大门，进一步拓展了分析的维度，卷积神经网络(CNN)能够轻松处理图像类商品数据，精准识别用户的视觉偏好；图神经网络(GNN)则可以深入分析用户的社交关系，探寻潜在消费影响路径。

2 数据科学驱动的电商用户消费行为分析模型构建

2.1 用户画像构建方法

用户画像构建，作为数据科学引领下电商用户消费行为分析的关键一环，其核心在于通过整合多维数据并提取关键特征，进而构建出能够精准代表用户行为的模型。基于数据科学的分类算法，如 K-means 聚类或 DBSCAN 密度聚类，可依据用户消费频次、客单价、品类偏好等维度，将用户划分为新用户、老用户、高价值用户及低价值用户四类[5]。高价值用户往往展现出较高的复购频率、较高的客单价以及跨品类购买的行为特点，而低价值用户则通常表现出较低的活跃程度，且对单一品类有着较强的依赖性。特征工程作为构建用户画像的核心技术手段，需要借助数据降维以及特征选择的方法来优化模型的整体性能。主成分分析(PCA)就像一个精明的数据侦探，能够从用户行为数据里精准地提取出核心维度，有效减少那些冗余信息的干扰；而随机森林算法能对特征重要性进行评估，精心筛选出对用户分类贡献最大的变量，例如由最近一次消费间隔(RFS)、消费频率(F)以及消费金额(M)共同构成的 RFM 模型。

2.2 消费行为预测模型构建

消费行为预测模型能够借助历史消费数据，深入挖掘出用户潜在的行为模式，进而为电商平台提供具有前瞻性的决策支持。借助时间序列分析技术的 ARIMA 模型，能够精准捕捉到用户消费行为中的周期性规律，就像能敏锐察觉到季节性促销期间用户消费热情高涨、消费额达到高峰的情景一样；而机器学习算法里的 XGBoost 或者 LightGBM，则擅长处理复杂的非线性关系，进而精准预测出用户未来会购买的品类、消费金额以及具体的时间节点。在模型验证的过程中，需要综合考量准确率、召回率以及 F1 值等多项关键指标，以此确保预测结果具备高度的可靠性。交叉验证技术就像是一位细心的守护者，能够有效避免模型陷入过拟合的困境，而集成学习方法则如同智慧的指挥家，通过巧妙组合多个弱学习器，显著提升模型的泛化能力。

2.3 个性化推荐系统实现

个性化推荐系统作为数据科学在电商领域中的一个极为典型的应用，其核心要义在于借助用户行为分析，达成“千人千面”这般精准的商品推荐。协同过滤算法主要是依据用户之间的相似性或者商品之间的相似性来给出推荐，用户 A 把商品 X 和 Y 都收入囊中了，而用户 B 只买了 X，这时候系统或许就会把 Y 推荐给用户 B；内容推荐算法呢，则是通过细细分析商品的属性，像品牌、价格、功能这些，再和用户的偏好做匹配，从而生成一份推荐列表。深度学习模型中的神经网络协同过滤(NCF)方法，能够深入挖掘用户与商品交互过程中潜藏的特征，进而有效提升推荐的精准度。推荐系统对于实时性的要求，使得它必须具备高效处理数据的能力。像 Apache Flink 这样的流式计算框架，能够实时地捕捉到用户的行为数据，并据此对推荐模型进行更新；而 A/B 测试技术，通过对比不同推荐策略所产生的效果，来对推荐算法进行优化。

2.4 用户行为分析在营销策略制定中的应用

用户行为分析就像是给电商平台营销策略制定装上了一个数据驱动的智慧引擎，提供了坚实的决策框架。在进行目标市场定位时，要充分结合用户画像里地域、年龄以及消费能力等多个维度的信息，精准识别出具有高潜力的用户群体。促销活动的设计，应当紧密依托于对用户行为的精准预测结果，进而巧妙优化活动的开展时机与呈现形式。电商平台借助预测模型，能够精准识别出用户对于折扣、满减、赠品等各类促销方式的敏感程度，进而定制出差异化的活动方案。例如，某平台通过分析用户历史参与促销的数据，发现“满 300 减 50”活动对高客单价用户吸引力最强，而“第二件半价”则更受低客单价用户青睐。

3 数据科学在电商用户消费行为分析中面临的

挑战与应对策略

3.1 数据隐私与安全问题的处理

在电商用户消费行为分析的领域里，数据隐私保护已然成为了一个核心且关键的伦理问题。依据 GDPR 以及《个人信息保护法》的相关规定，电商平台要运用差分隐私(Differential Privacy)技术，对用户的行为数据开展脱敏处理工作，以此保证在分析流程里原始数据能实现不可逆加密。例如，在用户浏览轨迹分析中，可采用 k-匿名化算法对 IP 地址、设备 ID 等敏感字段进行泛化处理，使单个用户行为无法被反向追踪。同时，联邦学习(Federated Learning)这种技术框架的运用，能让数据处于“可用却不可见”的状态，它借助在本地进行模型训练以及聚合参数的办法，有效规避了因原始数据集中存储而引发的泄露风险。在安全防护的层面上，必须要构建起一个多层级且严密的防御体系。电商平台需以零信任架构(Zero Trust Architecture)作为基础，对数据访问开展动态权限控制工作，同时结合行为基线分析(Behavioral Baseline Analysis)技术，对异常访问行为进行实时监测。技术实施中需平衡隐私保护与数据效用。在技术的具体实施过程中，加密算法的挑选会直接对分析结果的精准程度产生影响，要是隐私预算(Privacy Budget)设定得过高，数据就会出现失真的情况，相反，要是预算设定得过低，又没办法满足分析方面的需求。

3.2 数据质量与偏差问题的解决

数据质量为消费行为分析提供了坚实的保障。在处理数据质量时，缺失值的处理是一个关键环节，需要紧密结合业务场景来挑选恰当的策略。具体而言，对于连续型变量，就像购买金额这类数据，我们可以采用多重插补法(Multiple Imputation)，依据协变量之间的关系来生成合理的估值；而对于分类变量，例如商品类别，则可以通过热卡填充的方式，从相似的用户样本中抽取合适的替代值。在检测异常值时利用孤立森林算法能够迅速找出离群点，此算法通过随机划分特征空间来构建决策树，那些异常样本因为路径较短，所以会被优先隔离出来，这种方法很适合处理高维且稀疏的电商行为数据。

样本偏差问题其根源往往在于非随机抽样的方式，分析结果就会自然而然地偏向那些高频消费群体。针对这种情况，可以运用分层抽样技术手段，依照用户活跃度、消费能力等不同维度来划分层次，保证各个层次中的样本比例和总体比例相契合。在特征工程这一环节里，要格外留意特征分布偏移状况，就像在节假日促销的时候，用户的购买频次会明显增多，这时可以运用时间加权的办法来调整特征的权重，让模型训练数据和预测期数据的分布能够相符。在算法层面，基于 XGBoost 的集成学习模型展现出独特优势，能够通过巧妙设置样本权重参数，有效降低高价值用户样本所带来的过度影响力，进而防止模型过度拟合少数群体的行为模式。

3.3 技术挑战与解决方案探讨

当面对如潮水般涌来的海量用户行为数据时，传统的机器学习算法，像 SVM、随机森林这类，或许会因为计算复杂度犹如高耸的山峰般过高，而难以做到实时响应。在这个时候，轻量级模型借助基于直方图的决策树构建方式以及梯度单边采样(GOSS)技术，能够明显减少内存的占用和训练所需的时间。在那些需要深入挖掘特征信息的场景中，卷积神经网络(CNN)和图神经网络(GNN)组合而成的混合架构，能够很好地捕捉到用户行为序列里隐藏的时空依赖关系，通过 GNN 搭建起用户与商品之间的交互图。计算能力的种种限制，能够借助分布式架构的有效运用来实现突破。Spark 生态系统里的 MLlib 库，能够为大规模数据的并行处理提供有力支持，它所依赖的基于 RDD 的内存计算模式，可以巧妙避开频繁的磁盘 IO 操作，进而提升模型训练的速度。当面对超大规模数据集时，可以运用参数服务器架构，达成模型参数在分布式环境下的存储与更新，就像在深度学习推荐模型里，借助异步梯度下降来协调多个工作节点，进而加快模型收敛的速度。此外，GPU 加速计算技术的巧妙运用，能够进一步压缩训练所需的时间周期，NVIDIA RAPIDS 库所提供的 cuDF 与 cuML 组件，可在 GPU 上达成与 CPU 端 Pandas、Scikit-learn 相兼容的数据处理以及机器学习功能。

4 结论

数据科学在电商领域中，扮演着用户消费行为分析核心技术支撑的角色，它借助用户画像的精心构建、消费行为预测模型的巧妙搭建、个性化推荐系统的精准推送，以及营销策略的细致优化等多元化应用，达成了对用户需求的深度洞察与精准回应。基于行为经济学理论与机器学习算法的融合，用户分群准确率与消费预测模型精度显著提升，如聚类分析可识别高价值用户群体特

征，时间序列分析则动态捕捉消费趋势变化，为库存优化与促销活动设计提供量化依据。展望未来，当联邦学习、边缘计算等前沿技术得到更深入的应用时，数据科学会持续发力，进一步引领电商行业朝着智能化、精细化的方向大步迈进。

参考文献

[1]李雨桐.电商平台大数据驱动的消费者行为分析与精准营销研究[J].经济与社会发展研究，2025(12):0063-0065.

[2]张翔，万鹏.大数据分析在消费行为预测中的应用研究[J].消费电子，2025(12):61-63.

[3]周丽梅，王春燕.电商平台用户行为数据挖掘与消费行为预测研究[J].老字号品牌营销，2024(9):18-20.

[4]覃贵申.电商主播对用户消费行为的影响分析——基于把关人理论的视角[J].电子商务评论，2024(2):3733-3738.

[5]杜松华，徐嘉泓，张德鹏，等.游戏化如何驱动电商用户绿色消费行为——基于蚂蚁森林的网络民族志研究[J].南开管理评论，2022(2):191-202.

作者简介：孙鹤诚，男，本科在读，研究方向：数据分析和治理、计算机应用和大数据分析。