目录

基于客户购买行为的跨境电商销售研究

摘要

本文以和鲸社区某部分美国电商消费者的数据集为研究对象，利用 Python 数据分析、数据挖掘和机器学习等方法，对数据集中美国客户的商品喜好、首选支付方式，运输方式偏好等购物行为数据进行多角度分析，从而探索跨境电商平台中影响客户购买意愿的因素。基于此研究，本文最终提出了一些对跨境电商平台商家的建议、策略，以及对文本未来研究的展望。

首先，引言提出本文的研究背景和研究价值，然后对数据集中客户画像进行具体分析。我们利用相关性热力图、条形图等数据可视化图像将复杂数据转化为直观图表和图像，结合美国本土的经济、地理、运输设施等因素，加以相关性分析和决策树模型，从而对客户商品、支付方式、运输方式等购买行为进行分析。研究结果显示，商品颜色、支付方式选择、运输方式选择等影响客户购买商品的因素对于提升跨境电商平台销售额及客户满意度中具有重要作用。

其次，对于此数据集中总销售金额最大的服装品类，为了研究其销售的各类因素中，采用了机器学习模型，如 XGBoost 回归分析、灰色关联度分析、支持向量机回归和层次聚类等方法，并结合美国各地区的气温和经济等状况，分别从不同角度探讨了影响客户购买服装的因素。

最终，我们能够对商家提出基于客户画像和消费行为分析的有效建议，帮助商家深入了解跨境电商平台中不同客户群体的需求和偏好，从而实现精准的市场定位和个性化营销，对于跨境电商平台有借鉴性意义。例如，通过将相似购买行为的客户聚类，有助于商家为其定制个性化商品和服务；对购买某类商品的客户所居城市进行聚类，提升购买频次较大的城市的广告推广和销售预算，有助于提升商品营业额。这些建议能够帮助商家从多角度决策售卖合适的商品并为客户提供更适合的服务，提升用户体验，最终推动跨境电商平台的持续增长和盈利。

关键词:跨境电商，客户画像，相关性分析，聚类分析，机器学习模型

Abstract

This paper focuses on the dataset of U.S. e-commerce consumers from the Hejing community, using methods such as Python data analysis, data mining, and machine learning to perform a multi-dimensional analysis of shopping behavior data, including product preferences, preferred payment methods, and shipping preferences. The goal is to explore the factors influencing customers' purchasing intentions on cross-border e-commerce platforms. Based on this research, the paper offers recommendations and strategies for cross-border e-commerce platform merchants and provides insights for future research.

First, the introduction presents the research background and value of the study, followed by a detailed analysis of customer profiles in the dataset. By using correlation heatmaps, bar charts, and other data visualization tools, we transform complex data into intuitive charts and images. In combination with local factors such as the U.S. economy, geography, and transportation infrastructure, correlation analysis and decision tree models are used to analyze customers' behaviors, including product preferences, payment method choices, and shipping preferences. The results show that factors such as product color, payment method selection, and shipping preferences play a significant role in improving sales and customer satisfaction on cross-border e-commerce platforms.

Next, for the largest clothing category in terms of total sales in this dataset, machine learning models such as XGBoost regression analysis, grey relational analysis, support vector machine regression, and hierarchical clustering were employed to study the various factors influencing its sales. By combining the temperature and economic conditions of different regions in the U.S., we analyze the factors influencing customers' clothing purchases from different perspectives.

Ultimately, we provide effective suggestions for merchants based on customer profiling and consumption behavior analysis. These suggestions help merchants better understand the demands and preferences of different customer groups on cross-border e-commerce platforms, enabling more accurate market positioning and personalized marketing. For example, clustering customers with similar purchasing behaviors can help merchants customize personalized products and services for them. Clustering cities where customers purchasing certain products reside can enhance advertising promotion and sales budgets in cities with higher purchase frequencies, which can boost product revenue. These recommendations assist merchants in making more informed decisions on selling suitable products and providing better services, improving the user experience, and ultimately driving the continued growth and profitability of cross-border e-commerce platforms.

Keywords: Cross-border e-commerce, Customer profiling, Correlation analysis, Cluster analysis, Machine learning models.

1. 引言

1.1 研究背景

1.1.1 跨境电商行业发展

近年来，我国政府出台了一系列支持跨境电商发展的政策文件，以促进外贸新业态的快速增长。2024 年 6 月，商务部等 9 个部门联合发布《关于拓展跨境电商出口推进海外仓建设的意见》，提出优化跨境电商出口监管、推进海外仓建设等举

措。2025 年 3 月，国务院常务会议审议通过新一轮跨境电商综合试验区扩围方案，强调推进通关、税务、外汇、数据流动等监管创新，以稳定外贸基本盘。此外，国家税务总局 2025 年发布公告，对跨境电商海外仓出口实行“离境即退税”政策，进一步降低企业资金压力。这些政策共同构成了我国跨境电商发展的制度保障，推动行业向规范化、国际化方向迈进。

近年来，跨境电商行业在全球范围内快速发展。随着全球化进程的加快和互联网技术的不断创新，跨境电商已成为全球零售市场中的关键力量。据统计，2024 年全球跨境电商市场的交易额已超过 6 万亿美元，并预计将持续增长。越来越多的消费者习惯于通过跨境电商平台购买海外商品，尤其是在中国、美国、欧洲等主要市场，跨境电商的渗透率较高，而其他新兴市场也在快速扩展。

图 1 近年全球电商零售额及增速

在跨境电商发展论坛上，专家指出，“B2C 跨境电商”模式是未来发展的最佳选择，因为它能够实现海外结算和增值，进而积极推动 GDP 增长。B2C 跨境电商是指中国政府推动的跨境电子商务项目或平台，旨在通过数字贸易和电商促进跨境贸易的增长。这种模式使得企业（通常是制造商或供应商）能够通过电商平台直接向消费者销售产品或服务。这一模式的特点包括:

1. 全球市场:卖家可以直接接触到全球消费者,突破国界的限制。
2. 直接销售:卖家与消费者之间通常没有中间商,降低了中间成本。
3. 跨境物流和支付:涉及跨国运输、关税、支付方式等多方面的挑战和机会。

1.1.2 研究问题提出

在全球化发展推动下，跨境电商已成为促进全球经济交流的重要途径。近年来，我国积极推动跨境电商政策，但在跨境电商 B2C 快速发展的过程中，仍面临以下问题：

①客户诉求的提升:消费者需求多样化和品牌差异化和定制化;
②客户留存下降:缺乏有效的客户维系策略和去同质化商品;
③支付问题:支付方式多样性不足;
④物流问题:物流选择多样化和时效性;
⑤文化差异:由于不同文化背景下导致顾客对商品偏好不同

尽管国内跨境电商行业迅速发展，但国内跨境电商平台多集中产品供应链、物流配送等环节的优化，而对国际客户的需求、行为及消费心理的深入研究相对较少。许多平台仍然依赖传统的粗放式营销策略，未能实现基于数据的智能化运营。随着全球电商竞争的加剧，如何通过大数据的客户画像研究来提升用户体验、提高国际客户忠诚度，已成为国内跨境电商亟需解决的重要课题。

1.2 研究意义

为了提升跨境电商平台的竞争力，平台商家需要加大关注客户购物画像，来有效精准地理解这些消费者的购物习惯、购物偏好，提升客户购买意愿。这对于跨境电商平台制定精准的市场策略和提升用户体验至关重要,也帮助商家优化库存管理、个性化营销，从而提升客户满意程度，提升商品市场竞争力。

2. 数据来源与处理

2.1 数据来源

本文数据来自和鲸社区,其中数据集由多个 GitHub 仓库和 Kaggle 数据集合并而成，主要关注部分美国消费者的购物行为、支付方式、购买偏好等方面，适用于零售分析、消费者行为研究以及购物趋势预测等领域。研究报告指出，2024 年中国对美国的出口货物贸易总额为 6882.8 亿美元，占中国出口总额的 14.6%[1]。双边贸易金额大，因此此数据集对分析跨境电商平台中美国客户乃至全球客户画像有可靠性和借鉴意义。

2.2 数据预处理

本文数据集需要进行无量纲化，有利于处理聚类计算和提高机器学习模型的训练精度；对于连续型变量，进行标准化或归一化，有助于提升模型的表现，尤其是当特征的量纲不同的时候。

① 处理缺失值和重复值:本文采用均值$\bar{x}$填充缺失值和去除重复值

```
print(f'查看缺失值:{data.isnull().sum()}')
```

```
查看缺失值:顾客ID          0
年龄              0
性别              0
购买的商品           0
类别              0
消费金额(美元)        0
地点              0
尺码              0
颜色              0
季节              0
商品评分            0
订阅状态            0
支付方式            0
运输方式            0
是否应用折扣          0
是否使用优惠码         0
购买记录            0
首选支付方式          0
购买频率            0
dtype: int64
```

```
print(f'查看重复值:{data.duplicated().sum()}')
```

```
查看重复值:0
```

图 2 数据初步检查

借助调用 Python 中 pandas 库里 isnull( )和 duplicated( )方法，初步检查未发现缺失值、重复值以及异常值。

②分类特征编码：后文使用的部分机器学习算法要求所有输入特征为数值型，因此我们采用标签编码将分类特征转换为数值型对于每个类别$c \in C$（其中C是所有类别的集合），我们为其分配一个唯一的整数标签:

$$Label(c) = i \quad 其中 \quad i \in \mathbb{Z}, i \in \{0,1,2,\dots,|C|-1\}$$

② Score 标准化：将每个特征的均值调整为 0，方差调整为 1，从而使所有特征具有相同的尺度。

$$x_{standardized} = \frac{x - \mu}{\sigma}$$

其中:

1. x是原始数据点
2. μ是均值

3. μ_k是第 k 个簇的质心
4. σ是标准差

④归一化：确保每个特征在同等尺度上进行处理。

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

3. 数据及客户画像分析

3.1 数据变量相关性分析

斯皮尔曼等级相关系数（Spearman's Rank Correlation Coefficient）是一种不依赖于数据分布的统计方法，用于衡量两个数据集之间的单调性关系[2]。

客户消费行为受许多因素，但通过聚焦关键变量，研究能够在可控范围内探讨核心问题，避免因涉及过多变量而使分析变得复杂。因此，本文将通过分析年龄、消费金额（美元）、商品评分和购买记录这四个常见的影响跨境电商客户影响行为的关键因素，利用斯皮尔曼相关系数探讨这些变量之间的相关性，评估不同因素对购买决策的影响。

斯皮尔曼相关系数的计算公式如下：

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

其中:

1. ρ: 斯皮尔曼相关系数,用来衡量两组数据之间的单调相关性。。

2. d_i第 i 对数据的秩差,定义为$d_i = R(x_i) - R(y_i)$,其中$R(x_i)$和$R(y_i)$分别是x_i和y_i的秩次。

3. n:数据点的总数。

4. $\sum_{i=1}^{n} d_i^2$:每一对数据的秩差的平方和。

调用 Python 的 Scipy 库将客户消费情况各变量之间可视化为斯皮尔曼相关系数热力图:

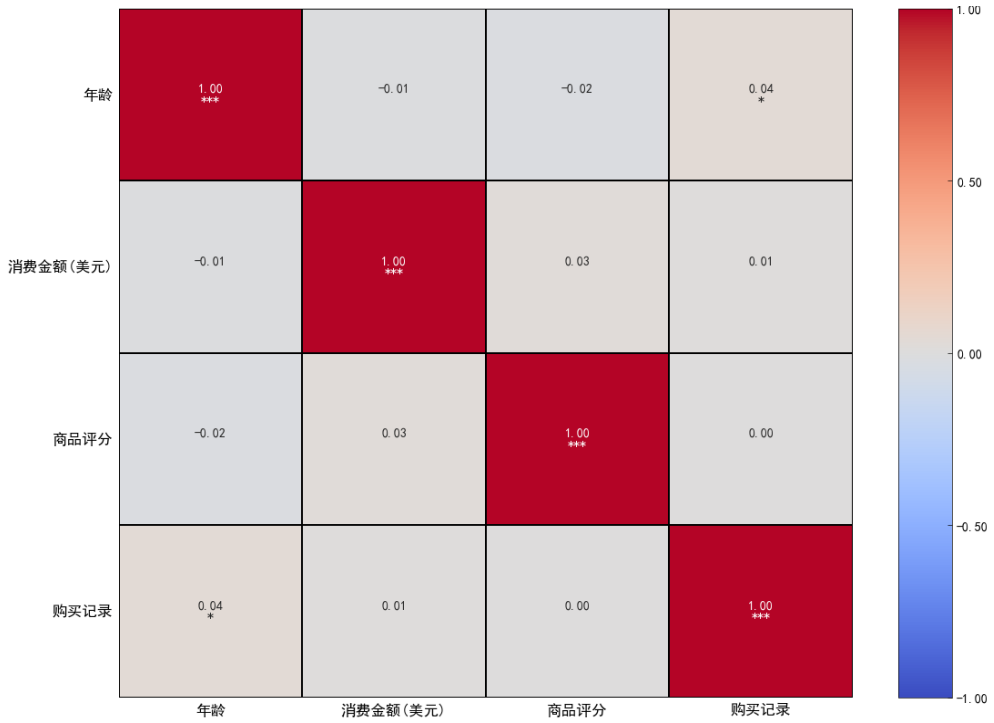

图 3 各变量之间斯皮尔曼相关系数热力图

由图 3 得知,年龄、消费金额、商品评分和购买记录相互之间斯皮尔曼相关系数均不大于 0.05，因此可认为不具有显著相关性，下文不再论述其之间联系。

3.2 商品喜好分析

3.2.1 商品购买量分析

词云图能够对各类的商品购买量进行比较，从而分出哪些是畅销品，哪些相对低销，**词云图**较大的字体表明了客户购买时主要集中在一些常见的商品上，而购买量较少的商品字体较小。

图 4 各商品购买量词云图

如上图,“裙子”“太阳眼镜”“女衬衫”等商品相对于其它较为畅销。

表 1 不同商品类别购买情况

类别	平均年龄	总购买记录数	平均评分
外套	44.3	8086	3.75
服装	43.7	43771	3.72
配饰	44.2	31900	3.77
鞋类	44.4	15114	3.79

基于《2024Q3 亚马逊 CPC 数据报告》和《2025 亚马逊卖家营销及品类数据报告》对跨境电商平台的商品营销分析，新品或未推广商品的自然曝光率极低，仅有约 5%-10%的消费者会浏览到第二页以后的商品，年均增长不足 10%[3]。然而，通过定向推广竞品 ASIN 或品类，可快速获取精准流量，平均 ROAS 可达 3:1 以上，头部卖家通过优化策略甚至达到 5:1，从而获得更高的利润[4]。

维度	未推广商品	推广商品
流量来源	自然搜索、偶然曝光（占比<10%）	付费广告、活动流量、精准定向（占比30%-50%）
转化率	普遍低于5%	可提升至10%-30%
长期稳定性	易受算法调整影响，波动大	排名稳定，用户粘性高
活动参与度	无法参与高流量活动	优先参与Prime Day、秒杀等

图 5 商品是否推广的影响对比

结合 2024-2025 年的已有最新数据与研究结论，亚马逊商品推广通过精准流量获取，流量曝光率达 30%以上，转化率提升至 10%-30%，显著拉开与未推广商品的销量差距。由此可得出商品推广对于销售很重要，我们推荐商家对销量较低的外套和鞋类商品进行推广，扩大其品牌知名度。

3.2.2 商品颜色偏好分析

通过颜色柱状图可显示不同商品类别在各个颜色偏好上的购买数量，每种颜色与商品类别柱形对应，帮助我们知道哪些颜色是客户在特定商品类别中的首选。

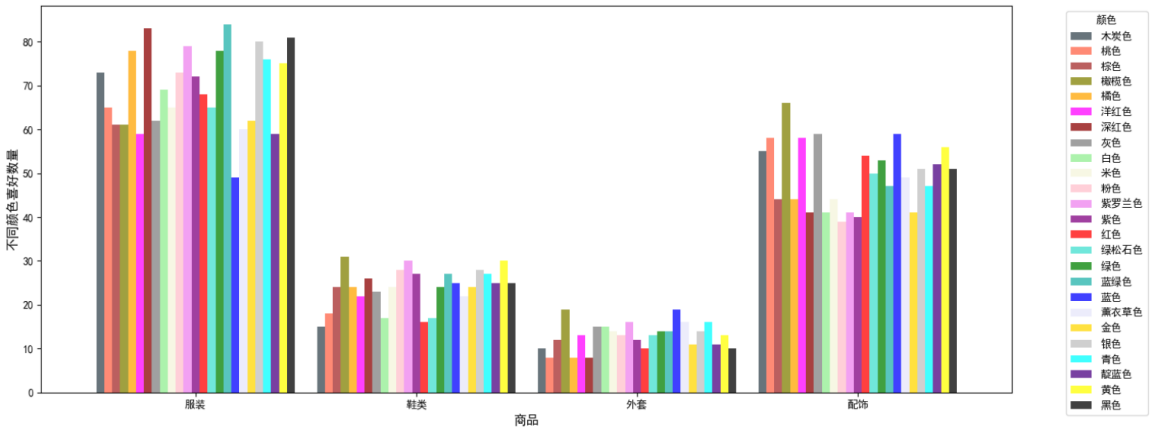

图 6 各商品颜色喜好人数柱状图

由上柱状图得出各类商品与其对应的首选颜色偏好：

表 2 各商品类别客户首选颜色偏好

类别	首选颜色偏好
服装	深红色，蓝绿色
鞋类	橄榄色，紫罗兰色，黄色
外套	橄榄色，蓝色
配饰	橄榄色，蓝色，灰色

与 RGB 模型（通过红、绿、蓝三色的组合表示颜色）不同，HSL 模型通过色相、饱和度和亮度三个参数来描述颜色，使得颜色的表达更贴近人类的视觉感知。我们以 HSL 亮度 50%为界限，亮度值低于 50%的颜色为深色，否则为浅色，从而探究客户颜色偏好。

1.亮度由下式计算:

$$L = \frac{\text{Max} + \text{Min}}{2}$$

- Max 是 RGB 颜色模型中 R、G、B 三个值中的最大值
- Min 是 RGB 颜色模型中 R、G、B 三个值中的最小值

2.计算色相:通过色轮的分段公式,得出具体角度

$$H = \begin{cases} 60^\circ \cdot \frac{(G-B)}{\text{Max}-\text{Min}}, & \text{if Max} = \text{R} \\ 120^\circ + 60^\circ \cdot \frac{(B-R)}{\text{Max}-\text{Min}}, & \text{if Max} = \text{G} \\ 240^\circ + 60^\circ \cdot \frac{(R-G)}{\text{Max}-\text{Min}}, & \text{if Max} = \text{B} \end{cases}$$

3.计算饱和度:

$$S = \frac{\text{Max} - \text{Min}}{\text{Max} + \text{Min}}$$

- 如果 Max 和 Min 的值相等(即颜色为灰色),饱和度为 0
- 否则饱和度越大，颜色的纯度越高

通过计算，得出每种颜色的亮度及深浅分类：

表 3 每种颜色亮度及深浅分类

类别	颜色及 HSL 亮度
深色	木炭色 0.261、紫色 0.251、橄榄色 0.251、深红色 0.273、绿色 0.251、蓝绿色 0.412、靛蓝色 0.255、黑色 0
浅色	桃色 0.639、紫罗兰色 0.722、洋红色 0.500、灰色 0.502、白色 1.00、米色 0.912、粉色 0.876、红色 0.500、蓝色 0.500、薰衣草色 0.941、金色 0.500、银色 0.753、青色 0.500、黄色 0.500、绿松石色 0.565、棕色 0.406、橘色 0.500

结合图和表，深色偏好(HSL<0.5)主要集中在正式的主装（例如，外套、服装），购买这些商品的客户更喜欢稳重、成熟的颜色。对于休闲、时尚的品类，如配饰、鞋类，客户更倾向于明亮、活泼的浅色(HSL>0.5)。

3.3 支付方式分析

在此数据集中，客户支付的方式有六种类型，分别是 Venmo(美国个人用户在线支付)、信用卡、借记卡、现金、贝宝(国际化支付平台)、银行转账。

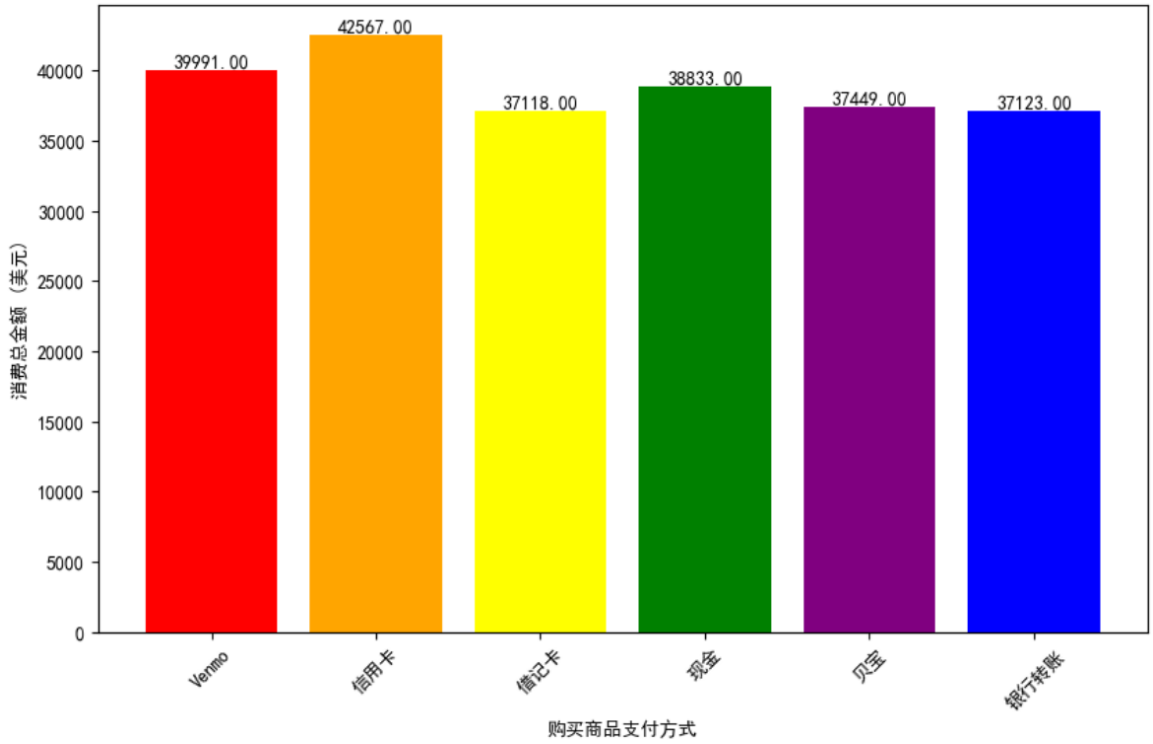

图 7 每种支付方式对应总金额

由上柱状图，这张图展示了客户不同支付方式对应的消费总金额。从图中可以看出，“Venmo”和“信用卡”分别是消费金额最高的两种支付方式,它们的消费金额明显高于其他支付方式。使用银行转账支付消费总金额最少。

首先，这种现象的一个原因可以从心理机制的角度进行分析。杨晨研究指出，信用卡等移动支付由于便捷性和现金流出的不可见性，对心理刺激性小并且购物预约感增强，平均消费金额也随着上升[5]。

其次，在美国，银行转账支付需要较长时间处理，尤其是在跨行或国际转账时，往往需要几天的等待时间，进行银行转账时，客户需要手动输入账号信息、密码，过程繁琐，有些客户会担心自己的信息安全泄漏[6]。相比于银行卡，Venmo、信用卡等即时移动支付方式步骤简洁，操作简单，具有便捷性，因此也是造成图表 6 的原因[7][8]。

表 4 各支付方式特点

支付方式	贝宝	信用卡	现金	借记卡	Venmo	银行转账
特点	安全，便捷，全球广泛使用	提供信用额度和奖励积分	携带不便，即使	不涉及信用贷款，	流行于年轻人，适	过程繁琐，支付时间较长

			不涉及隐私	直接，相对安全	合快速小额支付	

购买商品时，客户期望首选支付方式与实际支付方式往往不一样。我们通过计算数据集中各客户的首选支付方式出现总次数,得出下表:

表 5 各首选支付方式期望使用人次

首选支付方式	贝宝	信用卡	现金	借记卡	Venmo	银行转账
期望使用人次	677	671	670	636	634	612

协方差（Covariance）是用于衡量两个变量之间关系强度和方向的一种统计量，表示的是两个随机变量在其变动过程中是否有共同的变化趋势。首选支付方式期望使用人次及其对应的消费总金额是连续变化的，依次可以用协方差衡量其相关性。

数学公式:

$$\mathrm{Cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$$

表 6 协方差数学公式符号说明

符号	意义
$X，Y$	两个变量
X_i,Y_i	第 i 个观测值
$\bar{X}，\bar{Y}$	分别为 X 和 Y 的均值
n	数据的数量

通过计算得出：首选支付方式期望使用人数与消费总金额的协方差为 17803.2，即存在正向的共同变化关系。**此表明随着各类支付方式期望使用人数的增加，消费总金额也有一定的增加趋势**。研究说明商家为客户提供合适的支付方

式，能够增加客户的购买意愿；而提供流行、便捷、即时、操作简单的支付方式，如贝宝、信用卡等移动支付方式，减少心理性消费刺激，更加提升客户购买欲望和预约感。

3.4 运输方式分析

基于相关研究，客户购买商品时可选择的运输方式对其购买决策有重要影响，结合实际生活，此数据集里主要影响客户运输方式的选择有三个因素：购买商品类别、客户所居地点、支付方式。为了探究这两个因素哪个对运输方式的选择影响大，我们先利用随机森林模型来预测“运输方式”，通过模型训练分别获得类别和地点重要性，辅以互信息分析进行验证。这样能够从不同角度判断是“品类”还是“地点”对“运输方式”的影响更大

①随机森林核心思路：

1. 从原始训练集 D（包含 N 个样本）中，采用有放回的方式随机抽取样本，生成 T 个大小为 N 的自助采样数据集：$\{D_1, D_2, \ldots, D_T\}$

2. 对于每个数据集D_t，训练一棵决策树。在每个节点进行特征划分时，从全部特征中**随机抽取**一个特征子集(大小为 m<M,其中 M 为总特指数)，再在该子集中选择最优特征进行分裂[17]。

3. 随机森林的最终输出是通过对所有决策树的预测结果进行投票得出的:

$$\hat{y} = \arg\max_{k} \sum_{t=1}^{T} I(\hat{y_t} = k)$$

- 其中$\hat{y_t}$表示第 t 棵树给出的预测类别，I为指示函数

4. 对于回归任务，将所有决策树的预测值取平均结果得到最终结果：

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} \hat{y_t}$$

③ 互信息是衡量两个随机变量之间相互依赖程度的指标，来自信息论的概念。对于离散变量 X 和 Y，互信息的定义为：

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y)\, log\, \frac{p(x,y)}{p(x)\, p(y)}$$

- p(x,y)表示$X = x$和$Y = y$同时发生的联合概率
- p(x)和 p(y)分别表示 X 和 Y 的边缘概率

互信息的值越大，说明两个变量之间的信息共享程度越高，也就意味着它们之间的依赖性更强。

最终计算结果如下：

表 7 各变量对运输方式的特征重要性和互信息

影响 / 对于 / 运输方式	随机森林特征重要性	互信息
购买商品类别	0.0575	0.0012
购买地点	0.8302	0.0290
支付方式	0.1122	0.0025

通过计算可知：相对于购买商品类别、购买地点，购买地点对运输方式的随机森林特征重要性和互信息最大，分别为 0.8302 和 0.0290，因此客户购买商品选择的运输方式受购买地点影响最大。以下通过热力图探究其两者联系：

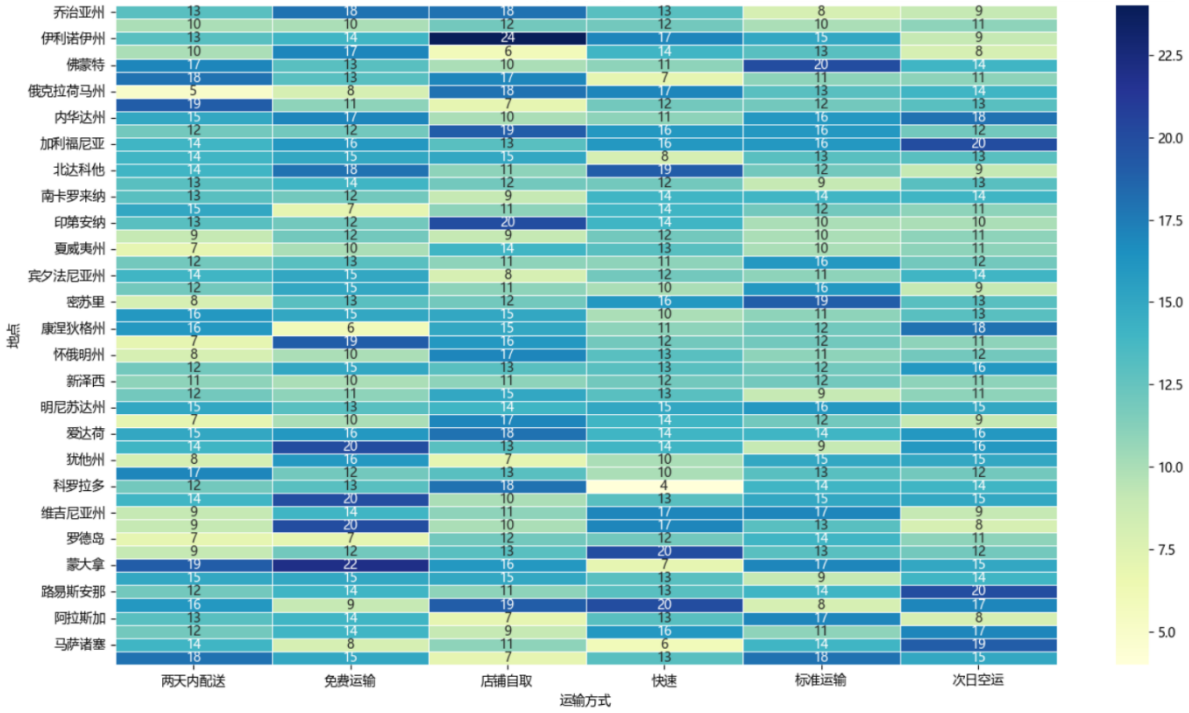

图 8 不同地点的运输方式选择热力图

对上述热力图进行分析：

①高低频运输方式：**免费运输**和**标准运输**是多数城市选择的主要方式，选择频率较高，显示为较深的颜色；**次日空运**和**店铺自取**的综合选择频率相对较低，颜色较浅。

②区域性偏好：像**纽约**、**加利福尼亚**等经济大城市，交通物流设施完善，商业活动繁忙，客户对快速配送（如次日空运或快速运输）需求高；而**阿拉斯加**、**蒙大拿**等偏远地区，由于物流基础设施不如大城市发达，运输成本较高，客户倾向于选择经济型的运输方式（如免费运输、标准运输），以减少运输费用。

艾媒数据中心的统计显示，美国服装采购渠道分布中，东部城市线下零售占比更高，而西部线上购物比例更高，归因于地理、经济、人口三个因素[9]。因此商家在制定不同地区的物流优化策略，需要考虑客户所居城市的 地理位置、经济水平、物流交通完善程度，更精准地提供合适的配送服务。

3.5 客户行为聚类

K-Means 聚类是一种常见的无监督学习算法，旨在将数据集划分为 K 个互不重叠的簇[10]。该算法通过最小化每个数据点到其所属簇的中心（质心）的距离，从而优化数据的分组。

对客户购物情况进行 K-Means **聚类分析**可以带来许多有用的洞察，帮助企业优化决策和提高效率,如个性化营销,库存管理和产品定制，发现潜在客户并进行行为预测。

K-Means 算法的目标函数(目标是最小化该目标函数)为:

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - \mu_k \|^2$$

其中:

- J：目标函数，表示所有数据点到各自簇质心的距离之和。
- C_k：表示第 k 个簇中所有的数据点。
- μ_k：是第 k 个簇的质心。
- $\|x_i - \mu_k\|^2$：是数据点x_i到簇中心μ_k的欧几里得距离的平方。

采用迭代优化的方式来最小化目标函数,步骤如下:

1.**初始化簇中心：** 随机选择 K 个数据点作为初始的簇中心。

2. **分配数据点到簇：** 对于每一个数据点x_i，将其分配给距离其最近的簇中心u_k，即将x_i分配给C_k，使得$\| x_i - \mu_k \|$最小。

3. **更新簇中心：**一旦所有数据点都分配到相应的簇中，重新计算每个簇的中心μ_k，即簇中所有数据点的均值：

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$

4. **重复步骤 2 和 3：**通过反复分配数据点并更新簇中心，直到簇中心不再显著变化，或者达到预定的最大迭代次数为止。

通过这些步骤，K-Means 算法能够逐步减少目标函数 J的值，直到收敛，达到局部最小值。最小化的过程是通过不断调整簇的划分和簇中心的更新来实现的。最终我们聚为 3 类，并用 PCA 降维展示:

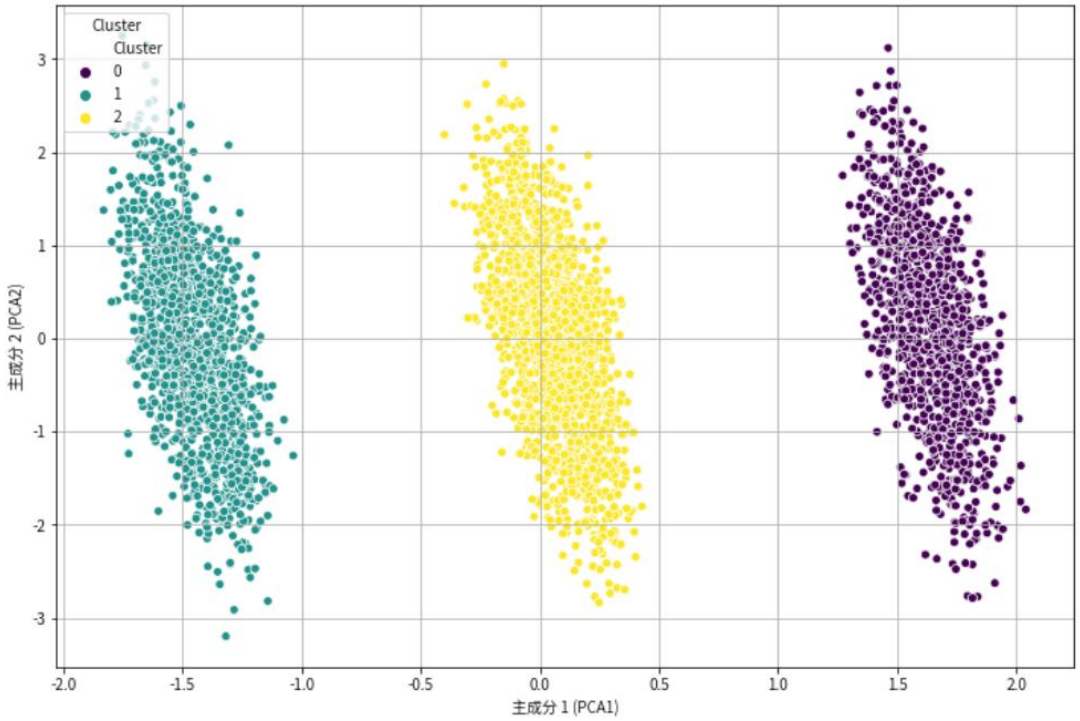

图 9 K-Means 聚类结果降维展示

聚类热力图是数据科学中一个非常有价值的工具，它通过可视化展示特征之间的相关性，帮助我们理解数据的结构和特征间的互动关系，进而支持特征选择、模型优化和业务决策。

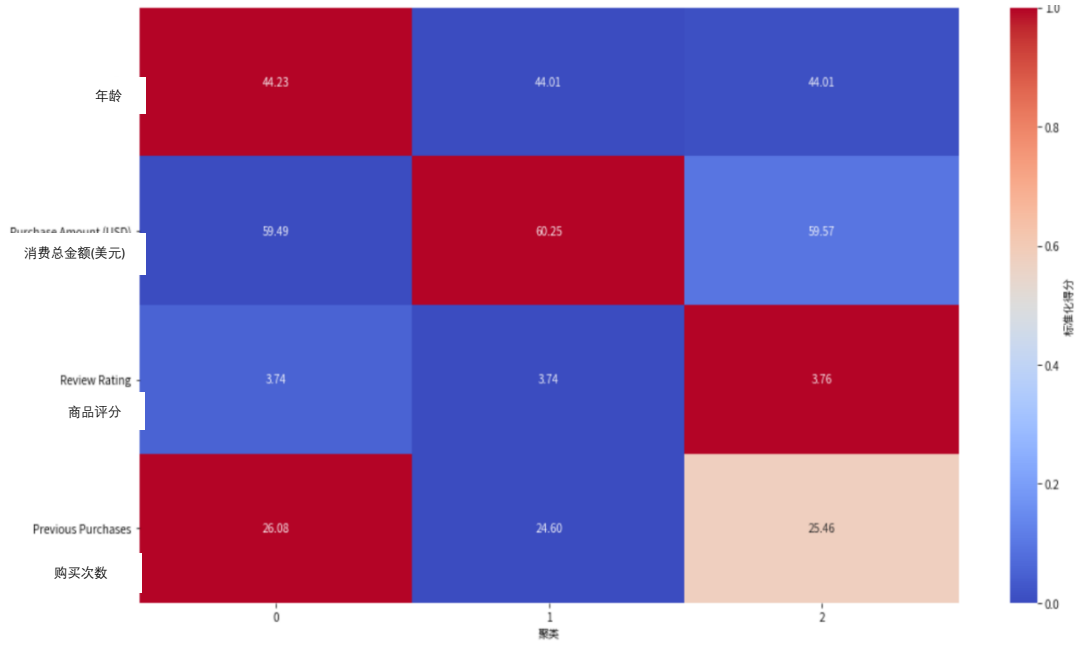

图 10 各聚类特征分布热力图

依据以上热力图得出如下表:

表 8 各变量的客户聚类情况

聚类 ID	年龄平均值(岁)	消费金额平均值(美元)	商品评分平均值	购买次数平均值(次)
0	44.23	59.49	3.74	26.08
1	44.01	60.25	3.74	24.60
2	44.01	59.57	3.76	25.46

分析上面表格,可得:

① 聚类 0:平均年龄最高（44.23 岁），够买次数平均值最多（26.08 次）。这类”熟客“对某些商品建立了忠诚度，对商品的折扣与优惠关注度高。

推荐商家针对此类客户措施:

- 提高忠诚度: 通过会员制度、积分奖励等方式增强客户的忠诚度。
- 定期推送新品和促销活动:提供专属的优惠券、折扣等。

② 聚类 1:平均消费金额最高（60.25 美元），购买次数平均值最少（24.60 次）。这类客户倾向于一次性大宗购买或选择高价值商品。

推荐商家针对此类客户措施:

• 提高用户粘性:通过建立更加人性化的客户关系，如主动联系、定期回访等，增加客户的购买频率。

• 定制化的促销活动:例如，针对大额消费提供专属折扣或赠品。

• 积分返利：每次购买高价值商品可积累积分，下次购买可用积分折扣或兑换奖励，从而鼓励顾客增加回购次数。

③ 聚类 2:商品评分平均值最高（3.76）。这类客户注重商品的质量和口碑，偏好购买评分高、口碑好的商品。

推荐商家针对此类客户措施:

• 加强商品评价管理:注重商品评价的质量,可以鼓励这类客户参与评价，给予奖励或优惠券。

• 质量保证:强调商品的品质和保障，向这类客户保证所有商品都经过严格筛选和质量检测。

4. 机器学习与建模

4.1 最大消费金额商品类别探究

我们运用 Python 中 Matplotlib 库和 Pandas 库，计算每个商品在每个类别下的平均销售金额和总金额并可视化:

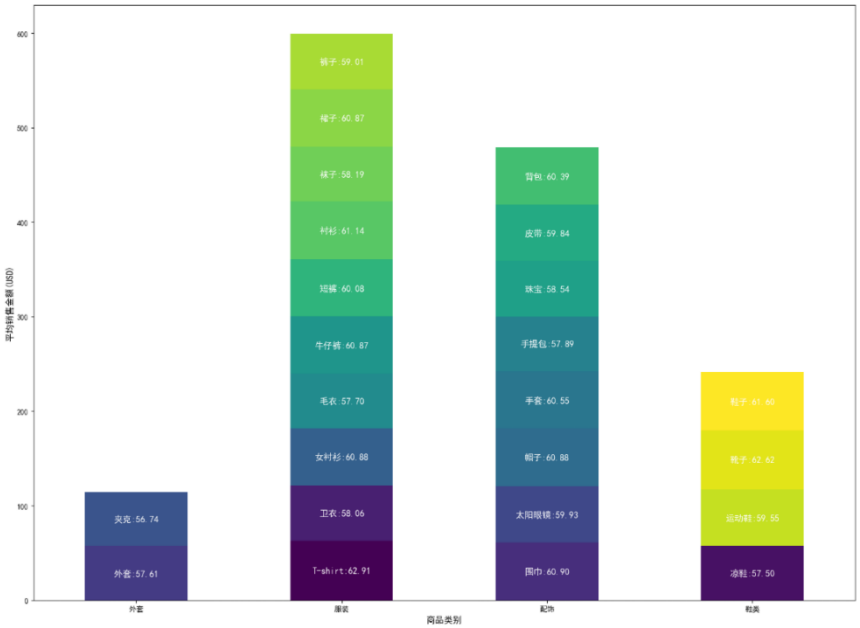

图 11 不同商品类别的平均销售金额图

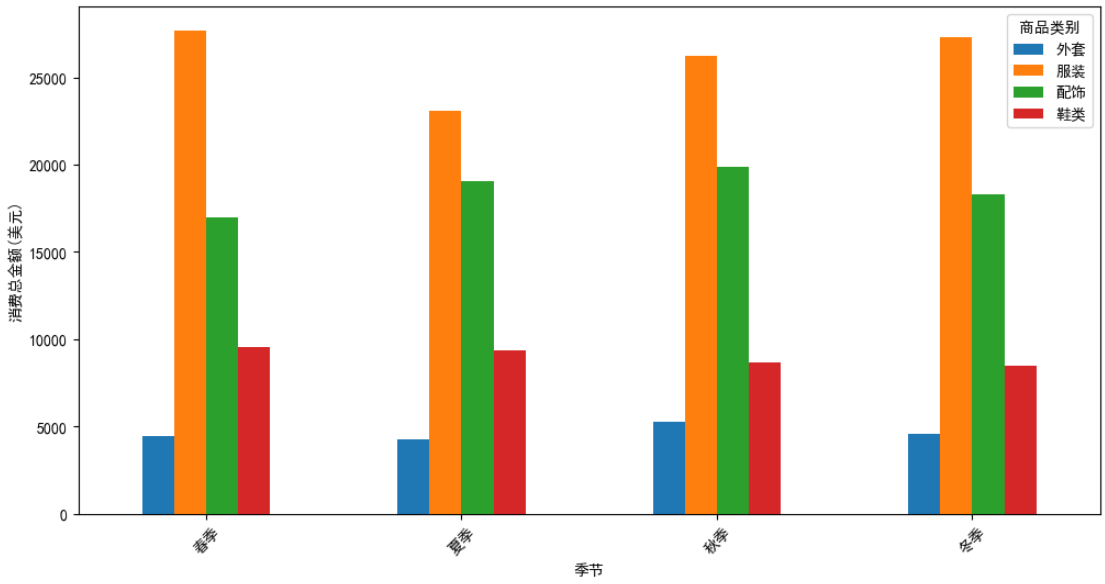

图 12 季节变化下的消费金额和商品类别

由上叠状图和条形图可知,服装带来的平均销售金额最大，其次是配饰，最后是外套。因此，以下我们研究影响服装销售金额的因素，从而帮助商家进行个性化决策，优化仓库管理。

```
季节
冬季    27274
夏季    23078
春季    27692
秋季    26220
Name: 消费金额(美元), dtype: int64
```

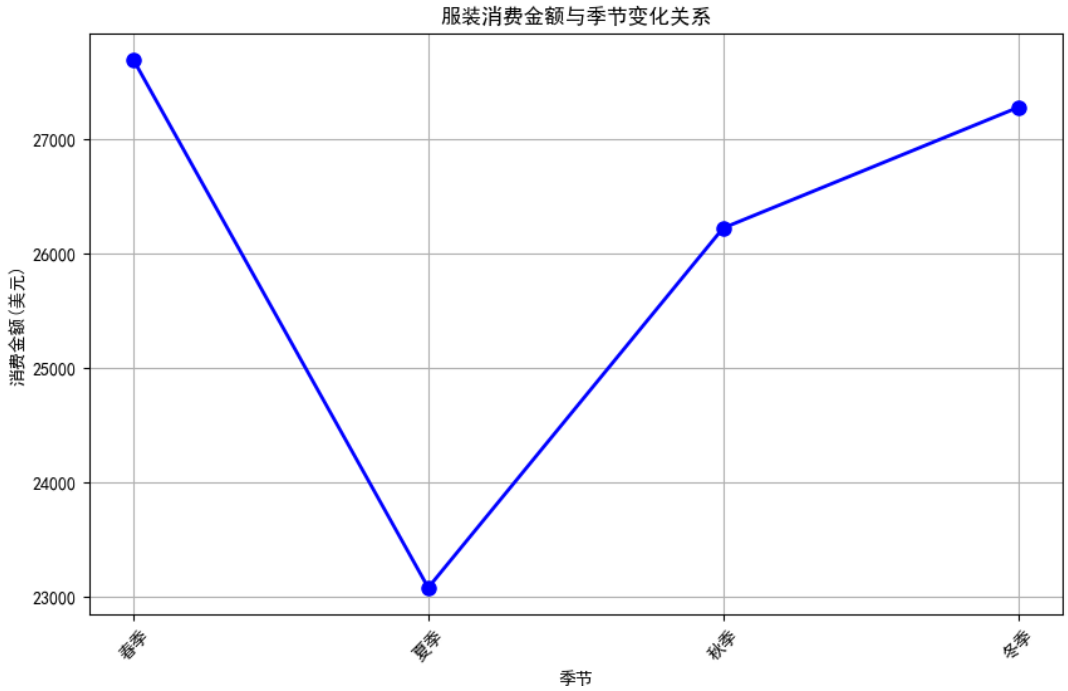

图 13 服装消费金额与季节变化关系图

由上条形图可得出结论，服装销售总金额与季节变化关系是先下降后升高。销售总金额峰值在春季（27692 美元），在夏季达到最低（23078 美元），随后在冬季上升至 27274 美元，结合实际生活，这销量值的变化往往源于季节性气温降低，添衣加暖的购买服装需求增大，购买保暖型服装客户人数增多。因此建议商家在**夏季减少保暖型服装储备量**，转型售卖凉爽型服装，减少过季商品和滞销商品的库存。

接下来我们将探究季节、运输方式和支付方式等特征变量对服装销售金额的影响，并确定哪个特征最重要。在 XGBoost 中，模型通过学习特征与目标变量之间的关系，并根据梯度提升算法逐步优化模型。

在此任务中，我们使用 **XGBoost 进行回归预测**，预测目标为季节、运输方式、支付方式对服装销售金额的特征重要性。XGBoost 的数学基础基于**梯度提升树**（Gradient Boosting Decision Trees, GBDT），每棵树的学习目标是最小化残差(预测值与实际值之间的差异)。

1. XGBoost 采用加法模型，将所有决策树的预测结果相加，以得出最终的预测值:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

2.XGBoost 的目标函数包括 **损失函数** 和 **正则化项**：

$$L(\Theta) = \sum_{i=1}^{n} \ell\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right)$$

3.正则化项通常为:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \parallel w_k \parallel^2$$

4.为了训练每棵树，XGBoost 利用梯度提升的思想，通过最小化损失函数来逐步改进模型。对于每个样本，我们计算**梯度**和**二阶梯度**：

$$g_i = \frac{\partial L(\hat{y}_i)}{\partial \hat{y}_i} \qquad h_i = \frac{\partial^2 L(\hat{y}_i)}{\partial \hat{y}_i^2}$$

5.然后,基于这些梯度信息,XGBoost 训练每棵树：

$$f_k(x) = -\frac{g_k}{h_k + \lambda}$$

上述公式的符号说明如下表所示:

表 9 XGBoost 数学符号说明

$\hat{y}_i$	第 i 个样本的预测值(服装销售金额)
K	树的总数
$f_k(x_i)$	第 k 棵树对样本x_i的预测
$\ell(y_i, \hat{y}_i)$	训练损失函数,衡量预测值与实际值之间的误差
$\Omega(f_k)$	第 k 棵树的正则化项
γ、λ	控制树复杂度的超参数

XGBoost 会根据训练结果计算每个特征对目标变量（如服装销售金额）的影响。在训练过程中，XGBoost 评估每个特征对损失函数的贡献，并通过特征重要性指标衡量其对模型的影响。**特征重要性越高，意味着该特征对预测贡献越大。**

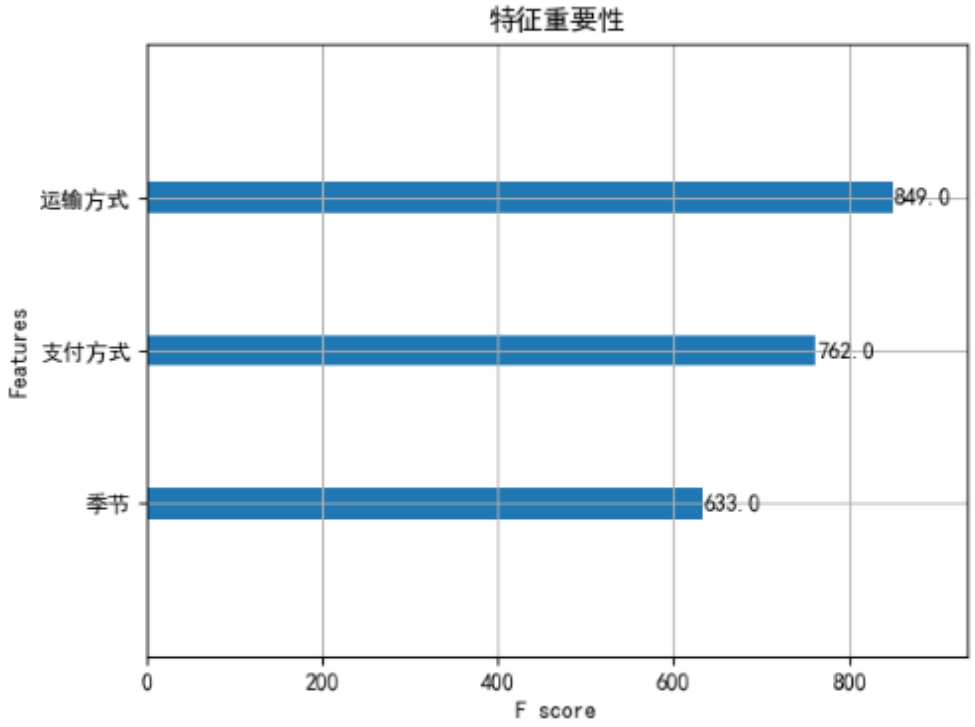

图 14 各特征对服装销售总金额的特征重要性

在 XGBoost 中，**训练次数**通常指的是模型训练过程中**树的数量**,即**迭代次数**（或称为 **boosting rounds**）。每一次迭代都会生成一棵树，XGBoost 通过多次迭代逐渐优化模型，减少训练误差。经过多轮迭代,得到上图。

从图 14 可以看出,在影响服装销售总金额的特征变量中，最为重要的是运输方式，特征重要性得分为 849.0,其次为支付方式和季节,得分分别为 762.0 和 633.0。因此，合适的运输方式对服装销售总金额有重要影响。有研究表明 ZARA 通过空运和快速专线配送，将服装从设计到上架周期缩短至 15 天，显著降低库存成本并提升销售额[11]，也同样说明合适高效的运输方式对产品销售十分重要。

上述使用的 XGBoost 模型有如下优点:(1)高效性;(2)模型复杂度较低；(3)适应性强，缺点如下:容易过拟合,不适合实时预测。

4.2 服装销售总金额与运输方式关联度

灰色关联度用于衡量两个序列之间的相似性或关联性。它是灰色系统理论中的核心概念，旨在通过比较多个系统间的相对变化趋势，揭示它们之间的关系强度[12]。灰色关联度越大，表示两个序列的变化趋势越接近，关系越紧密。以下计算和探究服装销售总金额与运输方式的灰色关联度。

灰色关联度计算包括以下几个关键步骤：

1.数据标准化:为了消除量纲影响，计算灰色关联度时通常需要对数据进行标准化。常见的标准化方法有极差标准化（将数据映射到[0, 1]区间）和 Z-Score 标准化（将数据转换为均值为 0、标准差为 1 的标准正态分布）等。

2.计算灰色关联系数(Grey Relational Coefficient, GRC)： 灰色关联系数是衡量两个序列在某个时刻的相似度，具体公式为：

$$\xi_i(k) = \frac{\Delta_{\min} + \rho\Delta_{\max}}{\Delta_i(k) + \rho\Delta_{\max}}$$

其中:

- $\Delta_i(k)$表示参考表示参考序列与比较序列在第k个点的绝对差异，公式为：$\Delta_i(k) = |X_0(k) - X_i(k)|$，$X_0(k)$是比较序列在第$k$个点的值,$X_i(k)$是比较序列在第 k 个点的值。

- Δ_{min}和Δ_{max}分别是所有差异中的最小值和最大值:

$$\Delta_{\min} = \min\left(\Delta_i(k)\right), \quad \Delta_{\max} = \max\left(\Delta_i(k)\right)$$

- ρ是分辨系数，一般取值为 0.5，用来调节计算中极差的影响。

3.计算灰色关联度（GRD）: 灰色关联度（GRD）是各个关联系数的平均值，通过计算每个时刻的灰色关联系数并取其平均值，得到总体关联度。

$$\xi_0 = \frac{1}{n}\sum_{k=1}^{n} \xi_i(k)$$

其中:

- ξ_0为灰色关联度，表示参考序列与比较序列之间的总关联度
- n 为数据的点数

灰色关联分析特别适用于处理数据较少或信息不完全的情况，可以从有限的信息中推测出系统之间的关系。根据计算结果，以下是每种运输方式与服装类别之间的灰色关联度：

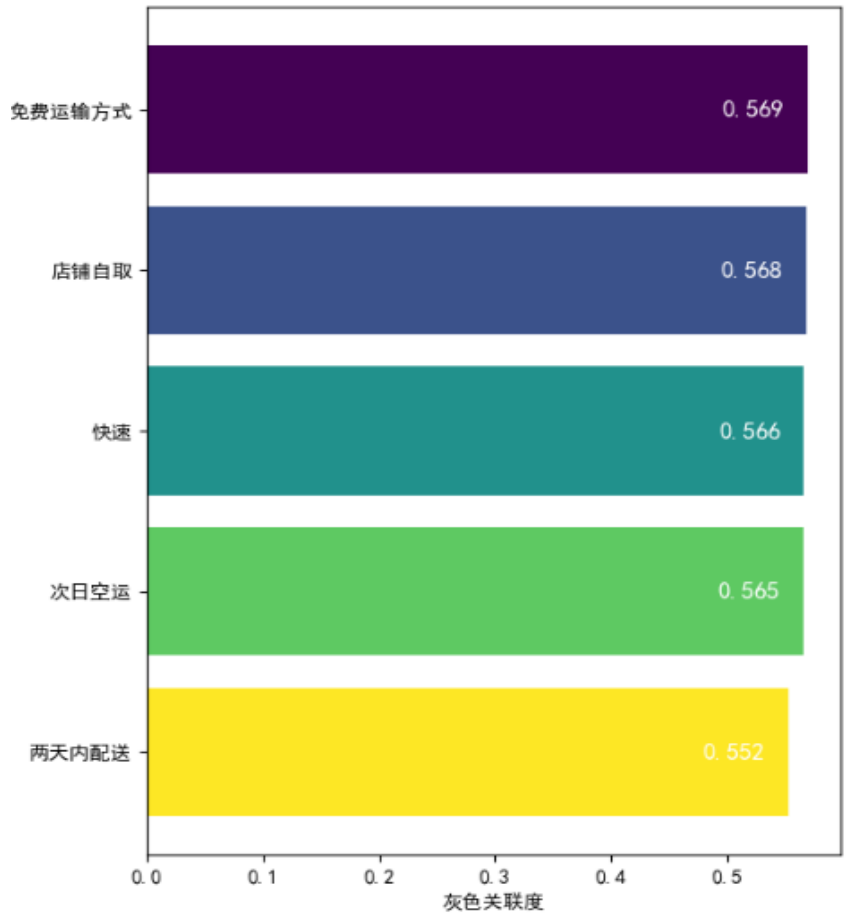

图 15 服装与不同运输方式之间的灰色关联度

由上漏斗图，不同运输方式之间的灰色关联度差值不大于 0.20(设定阈值)，可认为不具有显著差异性，因此**购买服装的客户对服装运输方式的选择并无太大差距**。但对于选择寄送服务（除了店铺自取以外的运输方式）的客户，免费运输方式与服装的灰色关联度最大，为 0.569，他们更希望免费运输，省去邮费；其次也有数量占比较大的客户更青睐店铺自取，与服装灰色关联度相比免费运输方式略小，为 0.568，他们选择亲自购买和查收服装，以此保证选择符合质量和心意的服装。

4.3 服装购买频率影响因素

支持向量机回归模型（SVR）是一种监督学习模型，广泛应用于分类和回归问题[13]。它实现对数据的分类的核心思想是通过找到一个最优的超平面来最大化类别之间的间隔。我们用此模型探究影响服装购买频率的不同因素重要性。

1.回归模型： 支持向量回归的基本目标是找到一个回归超平面，使得大多数数据点尽可能接近这个超平面。对于一个给定的输入x和输出y,回归函数$f(x)$为:

$$f(x) = \boldsymbol{w^T x} + b$$

•其中 w 是权重向量,b 是偏置项,x 是输入特征。

2.SVR 的优化目标：支持向量回归的目标是最小化目标函数，保证大部分样本的预测误差在一个给定的范围内（称为容忍度 ϵ）。因此，SVR 的目标函数是：

$$min_{w,b} \frac{1}{2}|w|^2$$

•约束条件:$|y_i - (w^T x_i + b)| \leq \epsilon$,对于所有的 i

•其中，ϵ是容忍误差的范围。此优化问题最终会生成一个最优的回归超平面。

3.核函数:对本数据集进行 MSE 和交叉验证，发现**线性核**表现略优于 **RBF 核**。这表明该数据集更适合使用线性核，因为其误差较小且更稳定。

$$K(x, x') = \mathbf{x}^T \mathbf{x}'$$

这里采用线性核函数将数据点直接映射到原始空间，不做额外的映射，模型以此找到一个最佳的回归超平面来预测目标变量（服装购买频率）。

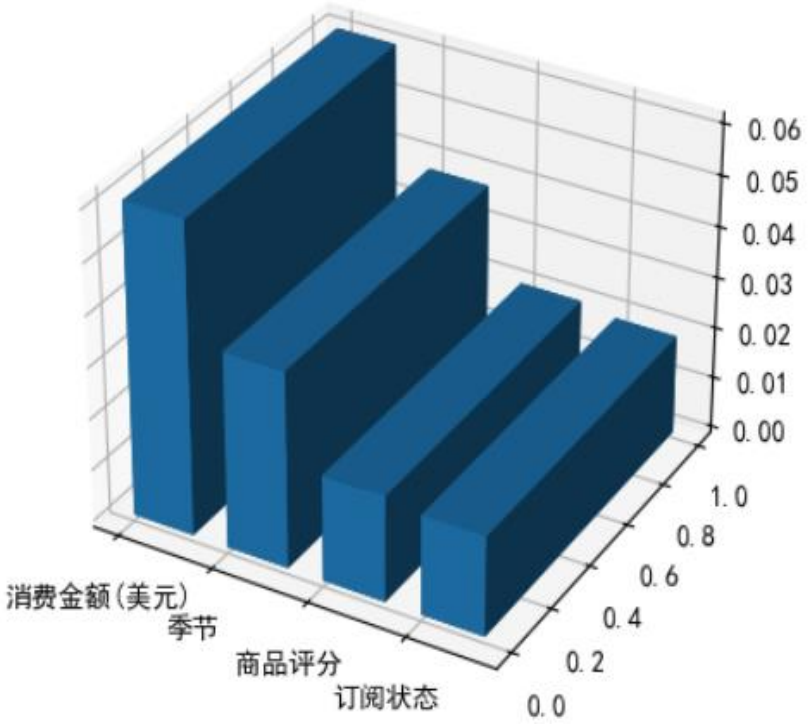

图 16 影响服装购买频率的不同因素重要程度

从图 16 可以看出，**服装购买频率受消费金额影响最大**，为 0.0608。据《中国纺织杂志》2022 年调查显示，34.03%的消费者将价格列为购买服装的首要因素，这一比例逐年上升，高价格敏感度导致消费者倾向于减少购买次数，或选择低价商品[14]；艾媒数据中心（2023 年）显示，消费金额与频次呈正相关，但高收入群体更倾向于低频高额消费，而低收入群体则高频低额[15]。此外，服装购买频率受消费金额影响还可从多角度进行分析，如价格敏感度、消费心理与替代行为转变等。

4.4 购买服装的客户聚类

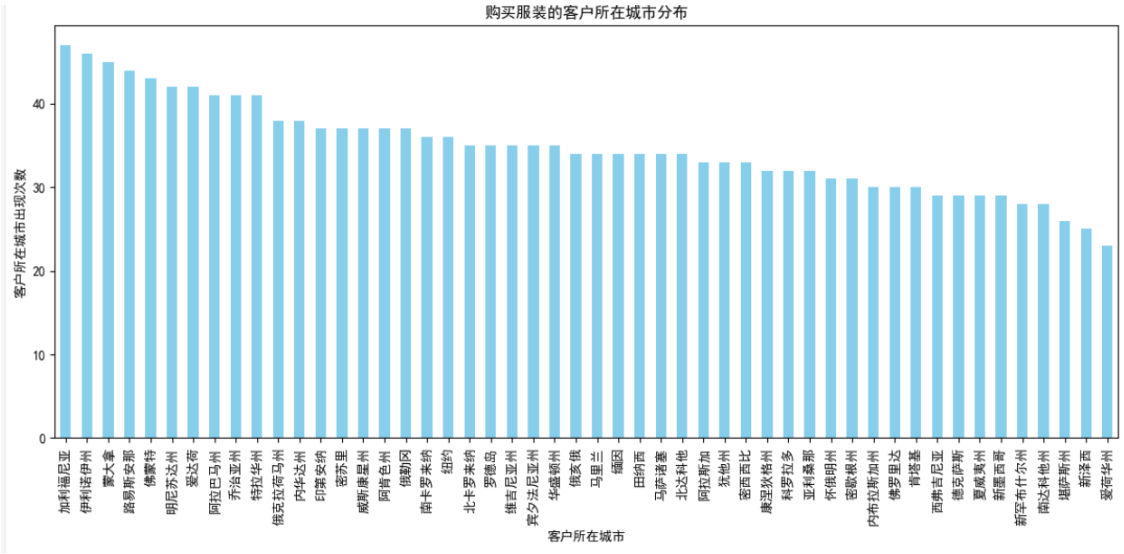

图 17 购买服装的客户所在城市分布图

在上图中,排在最前面的城市条形高度高，购买服装的客户较多，表明这些城市的消费者群体更大，或是市场需求更强烈，例如加**利福尼亚购买服装的客户最多**，为 47 人次。随着条形图向右延伸，城市购买服装次数逐渐减少，**爱荷华州购买服装人数最少**，为 25 人次。

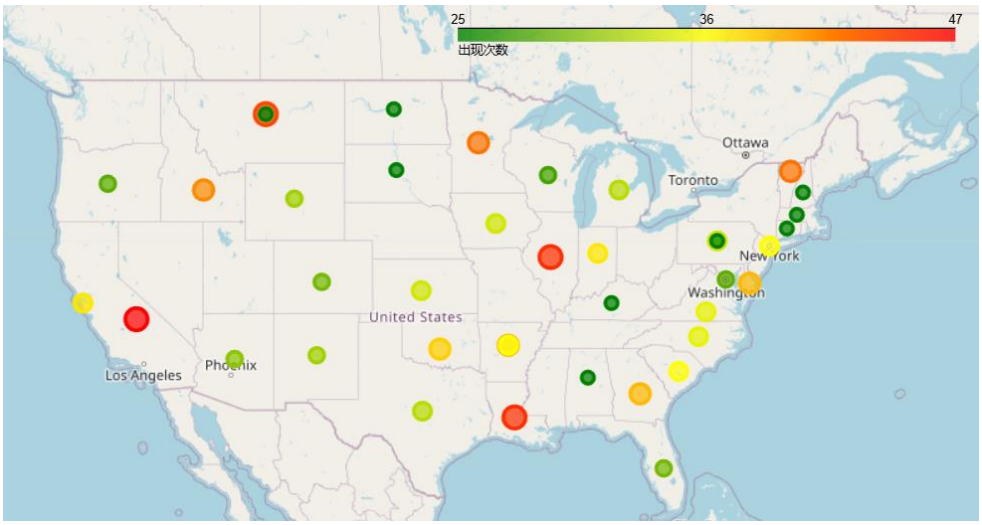

图 18 购买服装的客户居住城市分布图

上图为购买服装的客户所居城市在全美的分布，**其中点的颜色越深代表该城市购买服装的客户人数越多**。从图中可知美国东部城市购买服装人数多，且分布密集于沿海地区，西部分布稀疏且跨距大。这可从美国的地理和经济情况解释：美国东部主要为平原地形，沿海城市经济发达；西部则多高峻山脉如落基山脉，西部地区城市经济较为落后。如下是辅证的研究文章：**《服装行业地域分析》**提到，北美服装市场主要集中在都市和购物中心区域，与东部沿海城市高度重合[16]；**星淘惠跨境**的研究指出，美国服装进口主要依赖亚洲供应链，而东部港口（如纽约港、萨凡纳港）是主要入境点，这是东部服装消费市场的活跃原因之一[17]。

根据查询，购买服装客户数量多的城市有以下几个特点:①美国的重要经济中心;②冬季漫长;③夏季潮湿;④人口众多。同样，在购买服装客户数量较少的城市，有如下特点:①夏季炎热;②经济相对落后;③相较于前类城市人口较少。以此我们推测城市的气温季节，经济情况，人口数量，是影响该城市购买服装的客户数量的重要原因，后文我们将根据**购买服装的客户人数对城市进行聚类**。

层次聚类的核心在于计算不同城市（或簇）之间的相似度，并基于这个相似度逐步合并或拆分簇。因此能够帮助我们理解这些城市的购买行为的相似与异同，并将具有相似购买数量的城市归为一组。

层次聚类关键数学公式如下:

1. 欧式距离:用于计算城市(簇)之间的相似度

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

2.Ward 法:进行簇的合并，通过最小化每次合并后的簇内误差平方和（SSW）来进行聚类

$$SSW = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left(x_j - \mu_i\right)^2$$

3.链接矩阵:表示在合并过程中每个簇之间的相似度及合并顺序

$$\left(i\ j\ d_{ij}\ n_i + n_j\right)$$

上述公式数学符号意义：

表 10 层次聚类数学符号说明

x_i、y_i	城市购买数量在第 i 个特征上的值
$d(x,y)$	两个城市之间的相似度
k	簇的数量
C_i	第 i 个簇
x_j	簇C_i中的第 j 个数据点
μ_i	簇C_i的均值
i、j	合并的两个簇的索引
d_{ij}	簇 i 和簇 j 之间的距离
$n_i + n_j$	合并后簇的大小

经过计算并绘画出层次聚类的树状图：

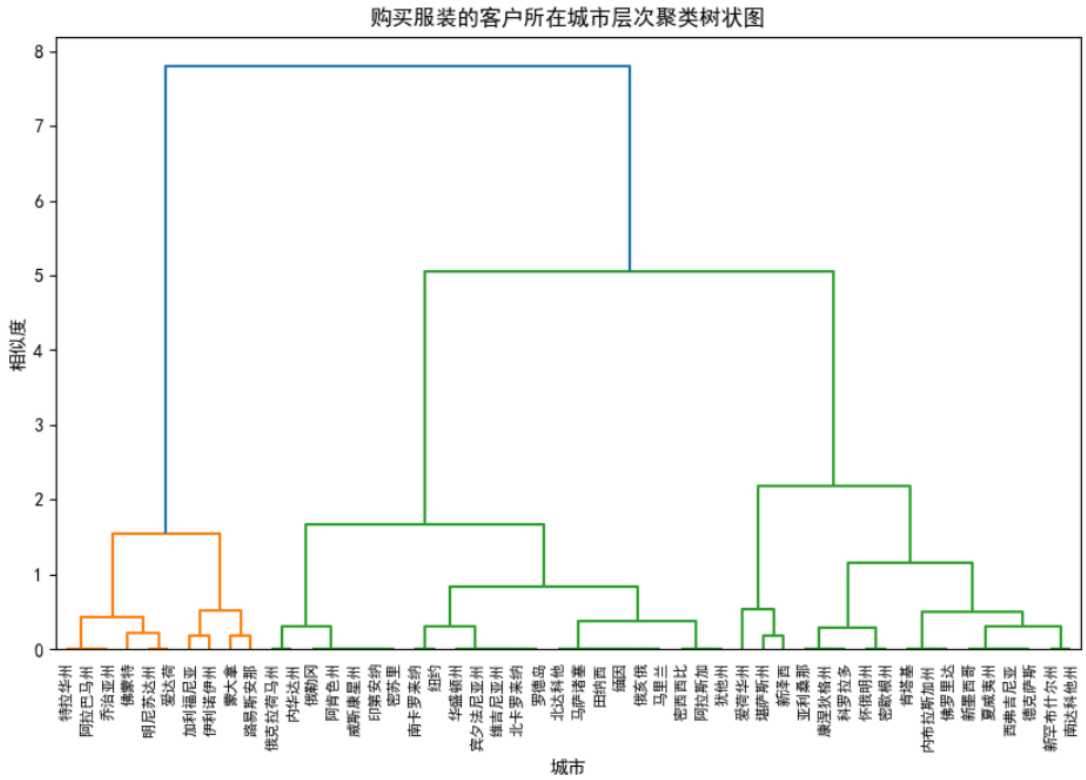

图 19 购买服装的客户所在城市层次聚类树状图

图 19 展示了购买服装的客户来自不同城市的数量的层次聚类结果，每个城市是树状图中的一个节点，城市之间的相似性通过“相似度”来衡量，相似度越高，连接的距离就越短。我们可以看出不同城市之间的购买服装的行为相似性，并将购买行为相似的城市分到同一组。**橙色区域中的城市购买服装的客户数量较多，而绿色区域城市购买服装的客户数量相对较少**。

因此对于跨境电商平台来说，**可以将更多的营销预算集中在这些高购买频次的橙色区域城市，将更多的广告预算分配在购买频次低的绿色区域城市**，优化市场推广策略，扩大服装销售利润率。

5. 建议与展望

通过本研究，我们深入分析了美国部分客户在跨境电商平台上的购物行为及其相关因素。研究发现，客户本身的购物心理、商品颜色偏好、支付方式选择和运输方式偏好等**内在因素**对其购买意愿具有重要影响，而客户所居城市状况等**外在因素**对其购买决策也有联系。Python **可视化、数据挖掘和数据分析**在文本对数据集的研究中，对分析客户画像发挥重要作用。此外，利用**机器学习方法**，我们探索

了不同的建模技术，包括决策树、XGBoost、灰色关联度分析、支持向量机和 K-means 聚类等机器学习模型和建模算法，进一步揭示了客户购买行为的复杂性。基于这些分析，本文将提出对商家的若干优化建议和本文未来研究的展望。

(1)对跨境电商商家:

①扩大销量较低的合适**商品推广**程度，能够增加销售机会和曝光率；
②对客户进行颜色喜好调查，性格年龄分析，**销售符合客户的颜色喜好**的商品；
③应根据不同消费者群体的首选支付方式和运输方式选择进行方案优化，综合客户的购物心理，所居城市的经济、地理位置、物流运输设施完善度，提供多样化支付方式，**简化支付步骤**，**采用合适商品运输方式**;
③售卖销售金额大的商品时，可根据城市的购买频次特点，对购买行为相似度大的城市进行聚类，结合不同城市的特点，**调整市场推广和广告预算分配**，减少低效投资。

(2)对本文:

①本研究仅使用了部分美国消费者的数据，未来将涉及更多国家和地区的消费者行为数据，从而优化现有的机器学习模型，提升预测的准确性;
②未来可将多种数据源结合进行分析，包括社交媒体、评论反馈等，多数据模态融合，以全面了解消费者的需求和偏好；
③未来可研究数据分析和人工智能技术相结合，进一步精细化用户画像，精准把握不同用户群体的需求并通过个性化推荐、定制化广告等方式，动态销售增长。

参考文献

[1] 金融界. (2025). 中国 2024 年对美国出口 5246.56 亿美元，同比升 4.9% [网页文章]. 搜狐. https://www.sohu.com/a/848349249_114984

[2]王晓燕,李美洲.浅谈等级相关系数与斯皮尔曼等级相关系数[J].广东轻工职业技术学院学报,2006,(04):26-27.

[3] Pacvue. (2024). 2024Q3 亚马逊&沃尔玛全球电商 CPC 数据报告 [数据报告]. 搜狐. https://www.sohu.com/a/818281447_121656383

[4] 雨果网. (2025). 重磅！2025 亚马逊卖家营销及品类数据报告 [行业报告]. https://www.cifnews.com/article/171574

[5]杨晨,王海忠,钟科,等.支付方式对产品偏好的影响研究[J].管理学报 ,2015,12(2):264-275.

[6] 港通智信. (2025). 美国支付方式全景解析：从现金到数字支付的全面指南 [网页文章]. https://www.fuwuhk.com/a/102777.html

[7] 未央网. (2021). Venmo:融合社交要素的美国知名移动支付服务商 [网页文章]. https://www.weiyangx.com/389810.html

[8] PYMNTS. (2025). Consumer preferences in digital payments: Venmo leads in adoption due to convenience [Industry report]. https://www.pymnts.com/

[9] 艾媒数据中心. (2025). 美国市场服装和鞋类典型采购渠道分布 [行业数据报告]. https://data.iimedia.cn/12995125/detail/13085412.html

[10]陈宝楼.K-Means 算法研究及在文本聚类中的应用[D].安徽大学,2013

[11]360 文档中心. (2020). UR Fashion Case Study [网页文章].服装行业物流特点分析 - 360 文档中心

[12]曹明霞.灰色关联分析模型及其应用的研究[D].南京航空航天大学,2007.

[13]范昕炜.支持向量机算法的研究及其应用[D].浙江大学,2003.

[14]中国纺织工业联合会上海办事处. (2022). 2021 年消费者抽样问卷调查报告:服装消费模式多样化 直播带货发展加速 [调查报告]. 中国纺织杂志. https://so.html5.qq.com/page/real/search_news?docid=70000021_611610a524815552&faker=1

[15] 艾媒数据中心. (2023). 2023 年中国服饰消费者与去年相比购买频次变化 [行业数据报告]. https://data.iimedia.cn/31026926/detail/49228102.html

[16]原创力文档. (2023). 服装行业地域分析 [行业报告]. Book118. https://data.iimedia.cn/31026926/detail/13085412.html

[17]王奕森,夏树涛.集成学习之随机森林算法综述[J].信息通信技术,2018,12(01):49-55.