南開大學
Nankai University

我想开间小店

— 基于 R 语言的电商平台订单数据分析

姓名：于沁涵 学号：1910227

摘 要

电子商务平台是一个为企业或个人提供网上交易洽谈的平台。目前越来越多的人群开始入驻电商平台，与商家在线上进行交易，这也使得各大电商平台不断发展，行业竞争也日益激烈。然而由于互联网与客户的“多变性”，选择合适的指标和思路对一些电商平台数据进行分析仍是一个亟待解决的问题。

本文针对上述问题，目标研究商品商业价值并找到高价值客户，对某电商平台订单的数据进行分析，利用 Kmeans 聚类评估商品的价值，并结合 RFM 指标建立 Logistic 回归模型和基于基尼系数的决策树模型，提出了一套筛选高价值客户的流程与方法。它可以通过电商平台的订单得到高价值客户，帮助电商平台更好的进行选品和运营。

研究结果表明：对于商品商业价值评估方面，基于 Kmeans 聚类，我们将商品分为 5 类，可以得到商品的畅销量、退货率、盈利能力都可以作为评估商品商业价值的指标，且可以很好的对商品进行聚类。对于筛选高价值客户方面，Logistic 回归模型的分类正确率在 80.83%，AUC 值为 0.916，最佳阈值为 0.098；基于基尼系数的决策树模型的分类正确率在 87.61%，AUC 值为 0.917，最佳阈值为 0.209。决策树模型的综合性能要略优于 Logistic 回归模型，但两个模型拟合效果较好，因此可以用于筛选高价值客户。此外，我们还探索了基于 RFM 指标的聚类分析，也可通过 Kmeans 聚类的方式对高价值客户进行筛选。

一、背景介绍

电子商务平台是一个为企业或个人提供网上交易洽谈的平台。企业、商家可充分利用电子商务平台提供的网络基础设施、支付平台、安全平台、管理平台等共享资源有效地、低成本地开展自己的商业活动。随着大数据、5G 时代的兴起，互联网技术越来越发达，拥有着快捷、智能化的突出优势，从前的人民币已经逐渐转化为线上支付，实体店也慢慢被各大如火如荼的购物平台占据。因此，越来越多的人群开始入驻电商平台，与商家在线上进行交易，这也使得各大电商平台不断发展，行业竞争也日益激烈。

对平台订单进行数据分析具有强烈的现实意义，它可以提高客户忠诚度、稳定客流，不断调整运营方向。这类分析工作往往专注于解决以下几个问题：与前一个季度/年度相比销售情况变化如何，如果下降、是什么原因导致的？用户消费行为有怎样的特征，高净值客户的行为是否有据可循？因此，越来越多的研究者开始投身此领域，企图对电商平台的客户做进一步分析，帮助电商平台稳定客流量，但是一些问题也应运而生。

近几年电商行业的各大网站纷纷通过降价、促销等方式来吸引用户，KPCB 的调查报告显示，2009 年到 2015 年全球移动端新用户的增长率持续下滑，可以预计在 2016 年这一增速将继续放缓。这意味人口增长带来的流量红利正在逐渐消退，用户增长将更加乏力，那么，通过单纯的价格战来吸引新用户的方式是否可行是一个值得思考的问题，可以看出寻找到合适的运营方式和数据驱动的经验、意识是必需的。因此，选择合适的指标和思路对数据进行分析仍是一个亟待解决的问题。

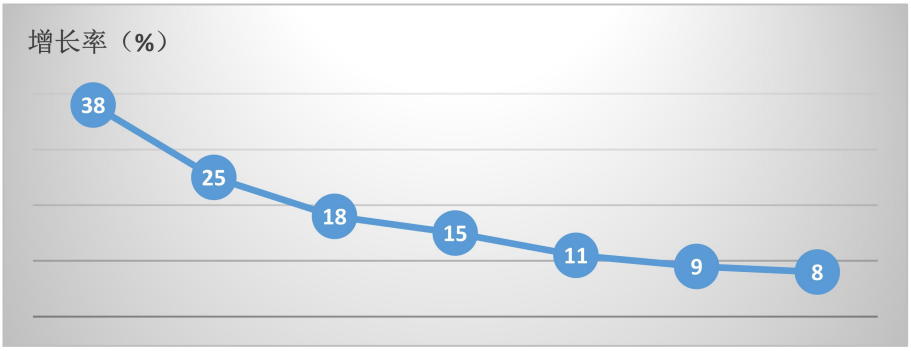

图 1　2009～2015 年全球移动端新用户增长率持续下滑

为了研究商品商业价值并找到高价值客户，本文针对电商平台订单的数据进行分析，提出了一套筛选高价值客户的流程与方法。它可以通过电商平台的订单得到高价值客户，帮助电商平台更好的进行选品和运营。

二、数据来源和说明

本文数据为某礼品批发电商平台一年内的订单。数据一共包含 8 个变量，共有 389168 个。其中，变量分别为发票号码、产品代码、产品描述、产品数量、时间、单价、顾客 ID、国家，具体变量说明如表 1:

表 1 数据变量说明

变量名称		详细说明	取值范围	备注
商品销售	发票号码	连续变量	536365～557214	每笔交易分配唯一的 6 位整数。退货订单的代码以字母'c'开头。
	产品代码	连续变量	10002～90208	每个不同的产品分配唯一的 5 位整数。
	产品描述	文本变量		对于产品的文字性描述
	产品数量	连续变量	-74215～74215	每笔交易的每件产品数量
	时间	连续变量	2010-12-01 08:26:00～2011-11-30 17:42:00	交易日期和时间
	单价	连续变量	0～38970	商品的单价（英镑）
客户情况	顾客 ID	连续变量	12346～18287	顾客在平台的 ID
	国家	分类变量	37 个国家	客户所在国家/地区

三、描述性分析

为了了解商品和客户的基本情况，我们首先进行简要的描述性分析，对数据进行初步了解。首先我们对商品进行分析。

（一）商品分析

首先我们来查看热销商品的销量情况，如图 2，可以看出电商平台在这一年内销售量最高的商品为 WORLD WAR 2 GILDES ASSTD DISIGNS，销量在 51852 次左右，而销量最低的商品为 JUMBO BAG STRAWBERRY，销量在 16257 次左右，可以看出，对于销量前 20 的商品，商品的销量还是有不小的差距的。下面我们继续探索这 20 个商品单价的情况，如图 3。

由图 3 可以看出，商品的单价与销售量走势相对是一致的，但是也有商品销量较差但是单价较高的商品，这些商品可能价格较高且实用性和质量不高，因此导致销量差。

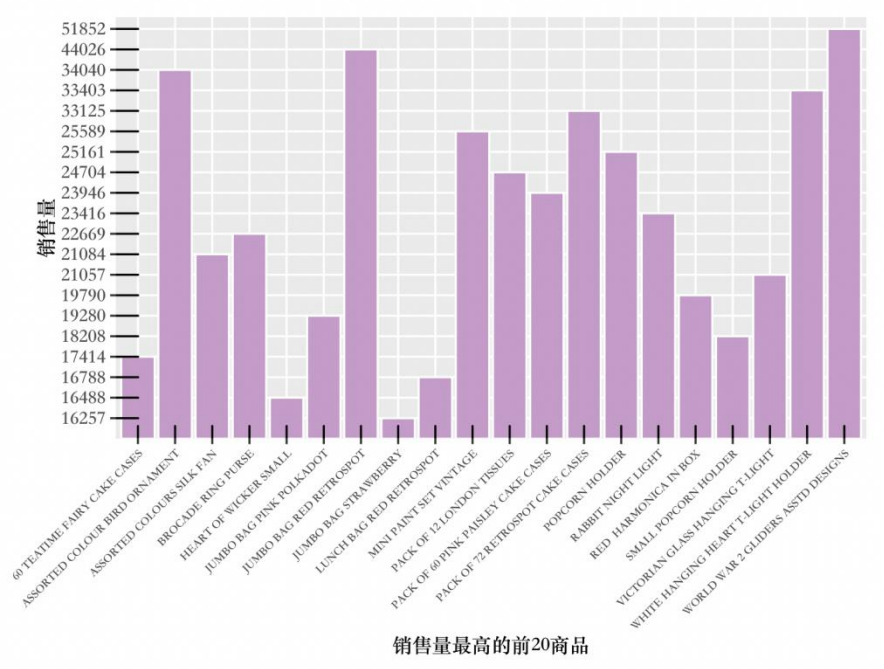

图 2 销售量最高的前 20 商品直方图

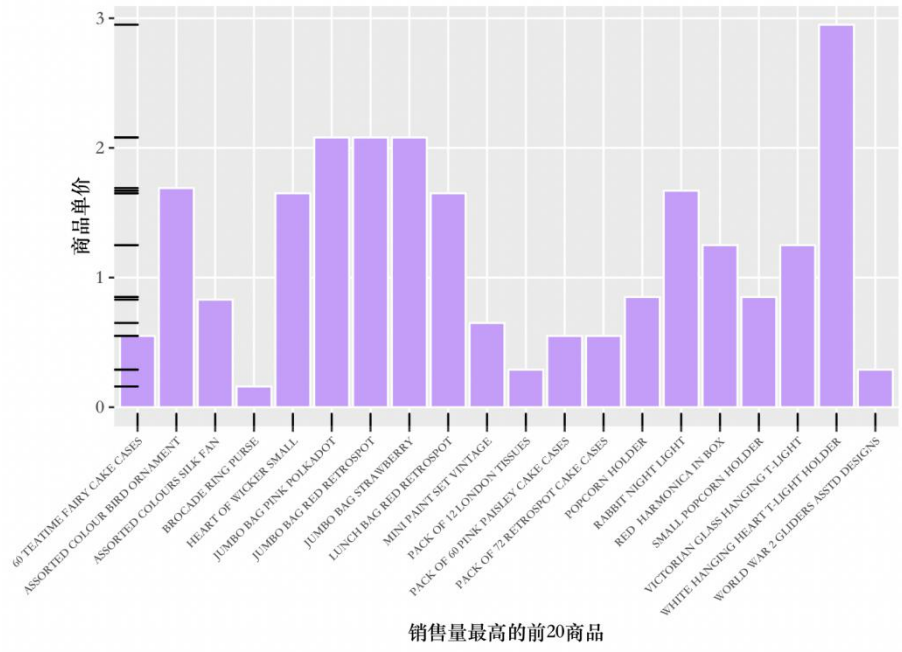

图 3 销售量最高的前 20 商品的单价直方图

接着，我们继续分析销量前 20 商品的“赚钱能力”。商品的销售额指的是商品的销量和商品单价的乘积，可以反应商品的“赚钱能力”。下图为销量前 20 商品的销量和销售额的排序直方图，可以看出，对于销售量前几名的商品，其销售额也排的较靠前，但同时也有一些黑马，例如销量排名最后的的 JUMBO BAG STRAWBERRY 的销售额排在了第二，这说明此商品非常有竞争力。

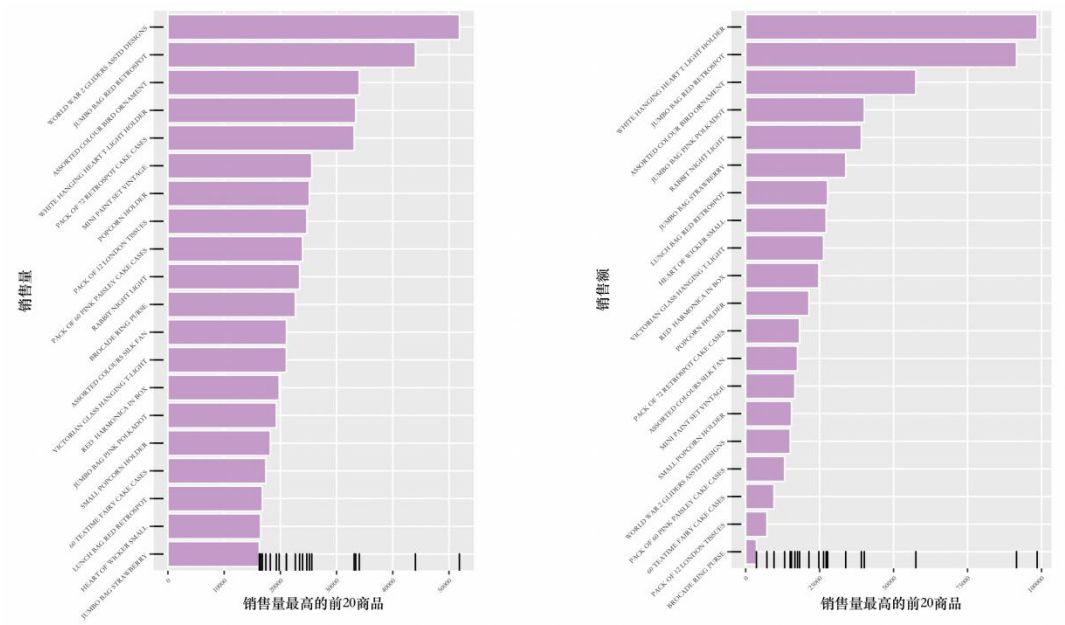

图 4 销售量最高的前 20 商品的销量和销售额直方图

（二）高价值客户行为分析

根据帕累托法则，即 20%的客户购买了 80%的产品/贡献了 80%的收入。企业中往往称这 20%能够推动公司 80%的收入的客户为“高价值客户”。在本文中，我们根据上述的帕累托法则定义高价值客户。因此，为了探究高价值客户在消费中的一些行为，我们简要做描述性分析。

1、高价值客户退货情况

首先我们查看客户的退货情况，我们绘制棘状图和箱线图来观察高价值客户和非高价值客户的退货情况，如图 5 和 6。从图中可以看出高价值客户的消费订单中含有退货订单的要高于非高价值客户，其退货率水平要高于非高价值客户。这是由于高价值客户对商品的要求更高，因此可能会出现退货情况。

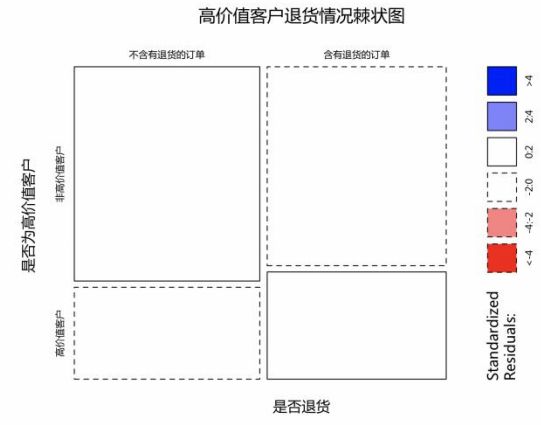

图 5 客户退货情况棘状图

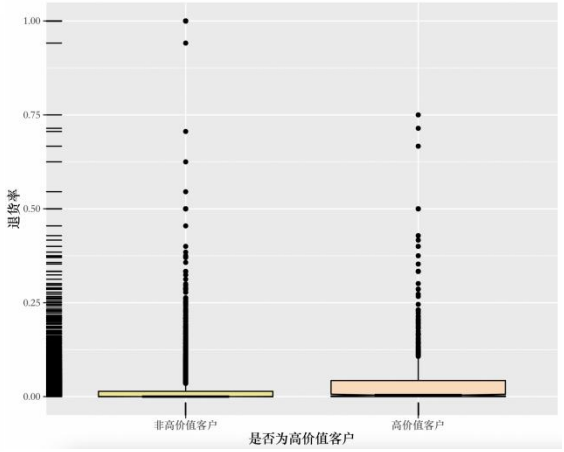

图 6 客户退货率箱线图

2、高价值客户消费情况

下面我们查看客户的消费情况，我们以客户总花费均值为阈值设置高低消费

水平，绘制棘状图，可以看出高价值客户会进行更高的消费，低价值客户则消费水平较低，这是因为高价值客户会购买更多的商品，与事实相符。

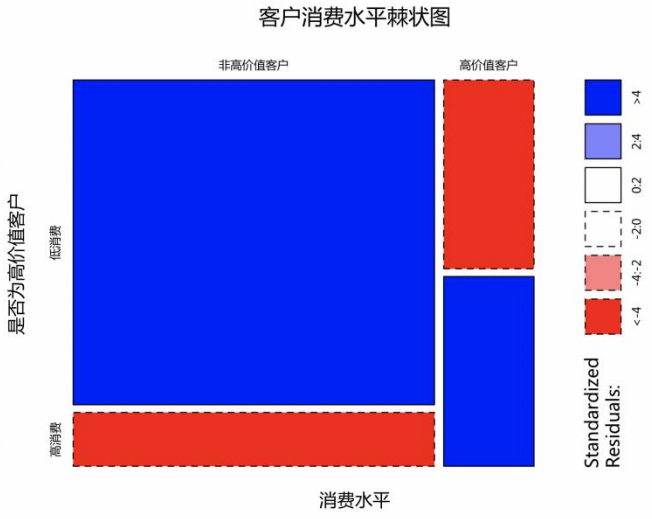

图 7 客户消费水平棘状图

从上述两个方面我们可以大致分析出商品和高价值客户的一些特征与概况，那么我们该如何对商品的商业价值进行评估？如何找到高价值客户并进行合理的预测？下面我们对这两个问题进行进一步的探究。

四、基于 Kmeans 聚类分析的商品商业价值评估

我们针对商品销售情况，设计出了一套评估商品价值的流程，基于 Kmeans 聚类的方法，建立相关指标，分别可以解释“是否畅销”、“是否具备高盈利能力”、“是否退货率较高”三个方面，以此来评估商品商业价值。

（一）特征提取：指标的构建

首先，我们进行指标的构建。我们建立了 3 个指标来衡量商品销售的情况，利用这 3 个指标可以进一步进行 Kmeans 聚类。分别为畅销量、盈利能力、退货率。具体解释如下：

- 畅销量：针对数据集中的商品统计该段时间每个商品一共售出的数量
- 盈利能力：利用销售量*商品单价得到每个商品在该段时间内赚的钱数
- 退货率：商品退货的数量/总交易数量得到每个商品在该段时间的退货率

（二）数据预处理

利用 R 语言软件，我们读入数据，发现没有缺失值和重复值，因此无需删除或填补数据。接下来，我们根据上述定义对指标进行构建，成立新的表格，并以 100*畅销量+100*盈利能力+100*退货率为得分对每个商品进行降序排序，取得分为正值的数据，一共 1621 行作为 Kmeans 聚类的数据，至此数据预处理结束。

（三）聚类类别的确定

接下来，我们对聚类类别进行确定。利用手肘法的肘部法则确定最佳聚类类别数：畸变程度的改善效果下降幅度最大的位置就是肘部，一般用畸变程度来确定最佳的值。

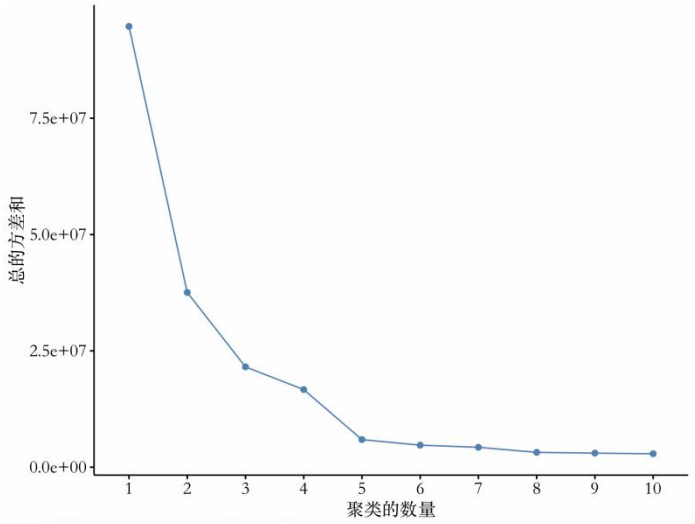

图 8 手肘法最佳聚类数量折线图

由图可以看出聚类类别为 5 的时候，聚类效果应该最好，因此在本文中，我们选择 K=5 的 Kmeans 聚类法。

（四）Kmeans 聚类

利用 factoextra 和 cluster 包，我们进行聚类，得到聚类相关结果图表如下。从聚类结果图中可以看出类别 3 的数量最少，类别 4 的数量最多，其中类别 3 的销售量、盈利能力较高、退货率较低，因此可以视为高价值的商品，但其类别较少，这说明在电商平台中，商家们仍然需要在提高商品价值上做出努力；类别 1 的销售量、盈利能力较高、退货率也偏高，说明此类商品比较常见，受广大人群喜爱，但可能会有与实物不符或者买到发现没有什么用途情况，故将其看作平价商品；类别 2 的销售量、盈利能力、退货率都适中、说明此类商品比较常见，在人群中算是普通的一类，故将其看作普通商品；类别 4 的各指标都偏低，说明此类商品比较小众，因此可以看作小众商品；类别 5 的销售量、盈利能力适中，但退货率非常高，说明此类商品的质量存在不小的问题，将此类商品看作风险产品。

表 2 聚类类别个数

类别 1	类别 2	类别 3	类别 4	类别 5
126	293	5	673	48

表 3 聚类中心

类别	销售量	盈利能力	退货率	得分
1	2.0508605	2.0681708	0.98634401	510.53753
2	0.9069249	0.8187366	0.38032901	210.59904
3	12.4536157	13.7299022	0.03687875	2622.03966
4	0.1563858	0.1386448	0.20919415	50.42247
5	3.2212846	2.7345639	4.24373297	1019.95814

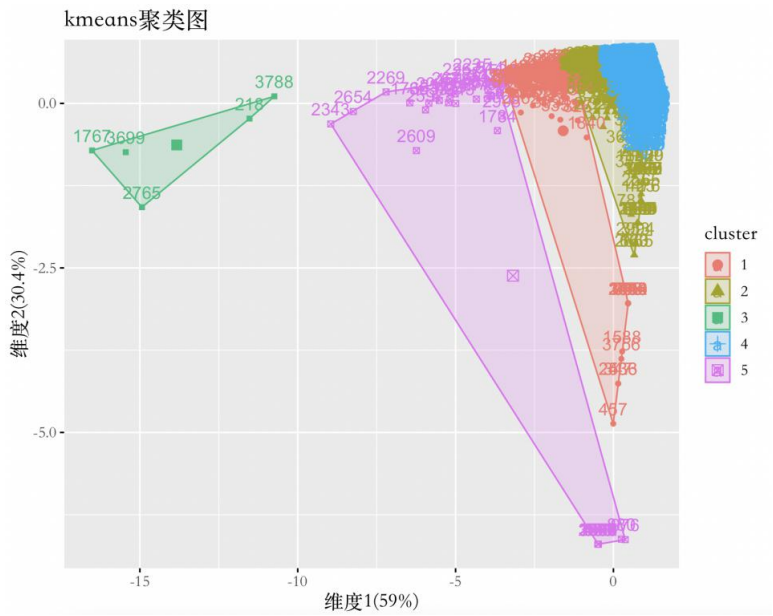

图 9 K=5 的聚类可视化图

五、基于 Logistic 回归模型的高价值客户挖掘算法

根据帕累托法则，即 20%的客户购买了 80%的产品/贡献了 80%的收入。企业中往往称这 20%能够推动公司 80%的收入的客户为“高价值客户”。

我们的目标是筛选出高价值客户，那么我们想到利用一些指标来对高价值客户进行建模预测，这样我们就可以通过这些指标进行拟合，来筛选出高价值客户。我们通过数据计算出了“高价值客户”，并记作 1，其余客户记作 0，将其作为因变量。因为因变量为 0-1 变量，我们很容易想到通过最传统简单的 Logistic 回归模型对高价值客户进行预测，希望找到筛选高价值客户的优良方法。因此为了了解高价值客户的行为，挖掘出最有价值的客户群体，我们将基于 Logistic 回归模型完成目标。

（一）指标的构建

我们首先综合 RFM 模型的指标维度以及平均活跃时长、购买量、消费金额、退单率设定模型所用到的自变量：

- 近度 R（Recency）：客户最近一次交易时间到当前时间的间隔，R 越大，表示客户越久为发生交易；R 越小，表示客户越近有交易发生
- 频度 F（Frequency）：客户在最近一段时间内交易的次数，F 越大，表示客户交易越频繁，F 越小，表示客户交易越少
- 强度 M（Monetary）：客户在最近一段时间内单次的平均交易金额，M 越大，表示客户价值越高，M 越小，表示客户价值越低
- 平均活跃时长：客户最早购买商品时间到最晚购买商品时间的平均时长，表明了客户使用电商平台的平均时间，可以反映客户的活跃程度，也就是平均客户生命周期
- 购买量：客户在该时段购买商品的总量
- 消费金额：客户在该时段消费的金钱数量
- 退单率：客户在该时段的退单率

（二）数据的预处理

利用 R 语言软件，rfm、plyr、lubridate 包对数据指标进行构建，将时间进行格式转化，构造出包含 R、F、M、平均活跃时长、购买量、消费金额、退单率的新表，对后四列数据进行标准化，至此数据预处理结束。

（三）Logistic 回归模型的建立

下面我们建立逻辑回归模型，其中因变量为 0-1 分类变量，1 表示是高价值客户，0 表示不是高价值客户，自变量包括标准化后的平均活跃时长、购买量、消费金额、退单率以及 0-1 变量 R、F、M，回归模型的结果如下表：

表 4 Logistic 回归结果（因变量：是否为高价值客户）

变量名称	标准化系数估计	标准差	统计量	P 值
截距项	-2.437236264	0.158355391	-15.39092697	1.88E-53
R	0.133413491	0.125480762	1.063218685	0.287682791
M	0.222047911	0.180847124	1.22782108	0.219514141
F	3.019951483	0.125489568	24.06535882	5.77E-128
平均活跃时长	0.15993493	0.049770485	3.213449304	0.001311509
购买量	0.119477528	0.435522667	0.274331367	0.783829982
消费能力	5.225264335	0.570074586	9.165931032	4.91E-20
退货率	0.196279765	0.047639542	4.120101833	3.79E-05

从回归结果可以看出，变量 F、平均活跃时长、消费能力、退货率都对是否为高价值客户有显著影响，我们又建立了 AIC 和 BIC 模型，发现两个模型都筛选掉了 R、M、购买量 3 个变量，因此我们只关注其中一个模型即可，AIC 模型结果如下。可以看出 AIC 模型的所有变量均显著，更适合进行预测，因此我们选用 AIC 回归模型。

表 5 AIC 逻辑回归结果（因变量：是否为高价值客户）

变量名称	标准化系数估计	标准差	统计量	P 值
截距项	-2.21913555	0.067732399	-32.76327981	1.96E-235
F	3.063399474	0.113229341	27.05482038	3.35E-161
平均活跃时长	0.173160966	0.049856558	3.473183297	0.000514324
消费能力	5.139175274	0.259094662	19.83512603	1.48E-87
退货率	0.19256478	0.047218024	4.078205003	4.54E-05

下面，我们对 AIC 准则下的 Logistic 回归模型的结果进行解读。我们发现，在控制其他变量不变的情况下，5 个自变量的标准化回归系数均为正数，这说明自变量与高价值客户呈正相关，也就意味着随着其他自变量的增多，是高价值客户的概率就会增大，这与指标的现实意义也相符。

因此我们可以总结出以下结果：频度越大，客户交易次数越多，是高价值客户的概率越大；消费能力越强，说明购买的东西越多，是高价值客户概率越大；退货率越高，说明对商品的苛求度越高，是高价值客户越大。

通过对模型结果的解读可以发现，回归结果和描述性分析的结果也一致。因此我们可以说模型的效果较好。

（四）模型预测与评估

为了进一步判断模型的正确性，我们接下来用测试集对模型进行预测，对比模型预测与真实的结果，计算混淆矩阵、绘制 ROC 曲线和覆盖率-捕获率曲线，对模型进行评判。

1、ROC 曲线、AUC 值与最佳阈值

对于逻辑回归模型，我们需要去寻找一个最佳阈值来将高价值客户组和与非高价值客户组进行更准确的划分，下图为测试集的 ROC 曲线，其中 ROC 曲线的弧度靠近左上角，模型的 AUG 值为 0.916，模型的拟合程度较好。预测的最佳阈值为 0.098，这说明当预测概率大于 0.098 的时候，预测为高价值客户，当预测概率小于 0.098 的时候，预测为非高价值客户。

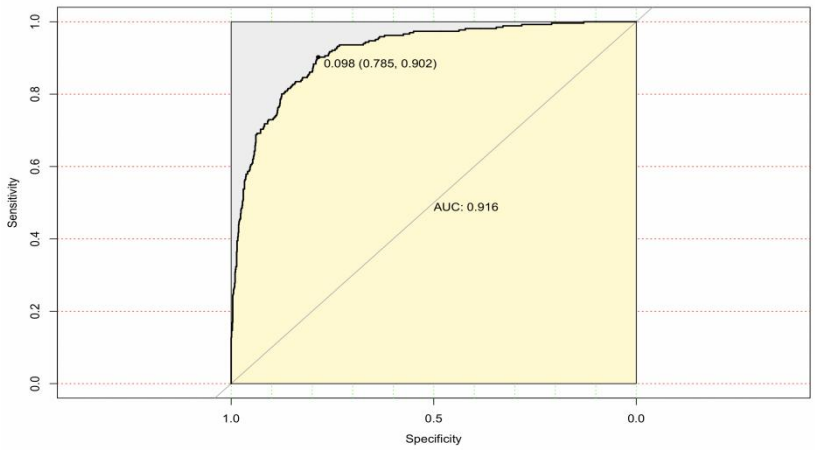

图 10 AIC 准则下 Logistic 模型预测的 ROC 曲线

2、混淆矩阵

对于逻辑回归模型，混淆矩阵用于比较分类结果和实际测得值，可以很好的反映分类结果的精度。不同的阈值也会产生不同的混淆矩阵，我们已经得到最佳阈值为 0.098，因此将最佳阈值带入模型可以得到混淆矩阵及相关结果，如表 6:

表 6 最佳阈值的混淆矩阵

		预测值		
		非高价值客户	高价值客户	总计
真实值	非高价值客户	810	26	836
	高价值客户	223	240	463
	总计	1033	266	1299

由表 6 可以看出整体的错判率为 (26+223)/1299=19.2%，TPR 为 240/463=52%，FPR 为 26/836=3.11%。

3、准确率

我们已经进行了最佳阈值选择，为了更直观的看出模型的好坏，我们对模型准确率进行计算，这是对模型好坏评估最直接的一种方法，最终得到结果为 80.83%，说明模型的预测能力较强，可以认为模型性能良好。

4、覆盖率-捕获率曲线

在实际中，我们还经常用成本收益曲线来度量的成本和收益。下图为 AIC 模型的覆盖率-捕获率曲线，我们发现其走势与 ROC 曲线类似，也反映了较好的结果。其中，在覆盖率为 20%的时候，捕获率可以达到 60%，说明覆盖预测概率最高的前 20%人，可以抓住 60%的高价值客户，说明模型精度较高。

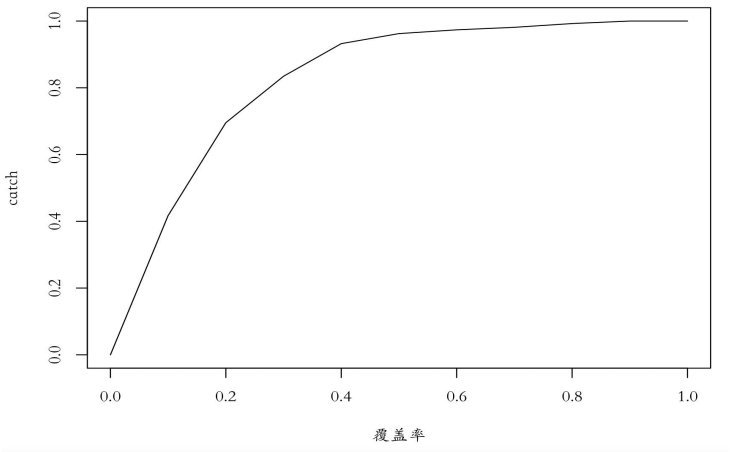

图 11 覆盖率-捕获率曲线

六、基于基尼系数决策树模型的高价值客户挖掘算法

决策树模型也是一个很好的分类算法，与 Logistic 回归模型同样的方法，我

们继续基于决策树模型进行建模，观察结果。在此我们不再展示指标和数据预处理部分，内容与 Logistic 回归模型相同。

（一）决策树模型的建立

首先，我们利用 rpart 包建立基于基尼系数的决策树模型，构建大的初始回归树，为了让树足够大，我们使用较小的 cp 值。这意味着指定较小的 cp 值，只要模型总体 R 方增加就继续产生新的分支。接着对回归树进行剪枝，使用 cp 值寻找最优值。最终得到剪枝后的决策树作为最终模型，可视化结果如图 12:

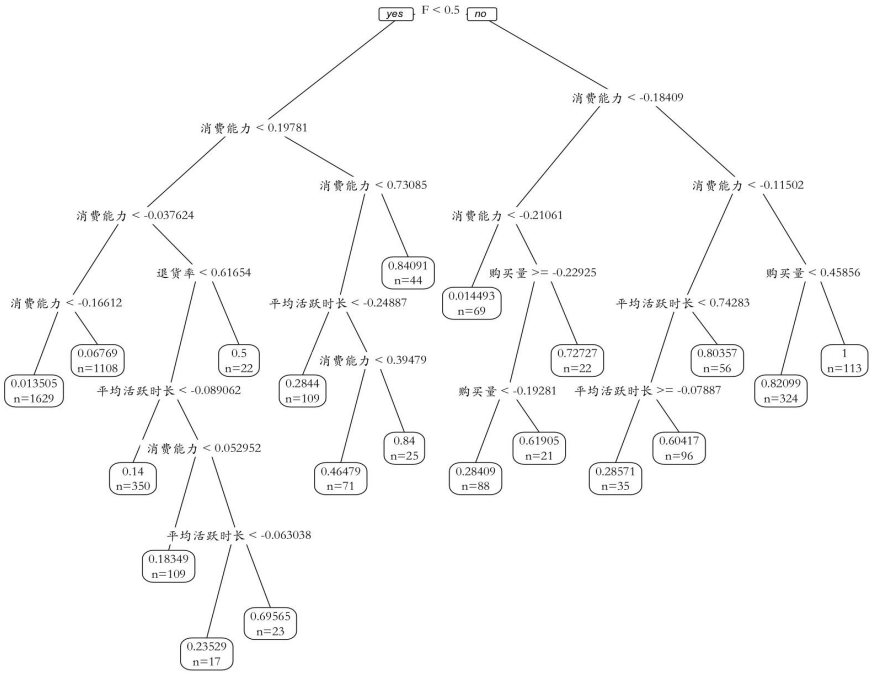

图 12　剪枝后的决策树模型可视化结果

（二）模型预测与评估

为了进一步判断模型的正确性，我们接下来用测试集对模型进行预测，对比模型预测与真实的结果，计算混淆矩阵、绘制 ROC 曲线和覆盖率-捕获率曲线，对模型进行评判。

1、ROC 曲线、AUC 值、最佳阈值与覆盖率-捕获率曲线

对于决策树模型，下图分别为测试集的 ROC 曲线和覆盖率-捕获率曲线，其中 ROC 曲线的弧度靠近左上角，模型的 AUC 值为 0.917，模型的拟合程度较好。预测的最佳阈值为 0.209，在覆盖率为 20%的时候，捕获率可以达到 60%，说明覆盖预测概率最高的前 20%人，可以抓住 60%的高价值客户，综上指标都可以说明模型精度较高。

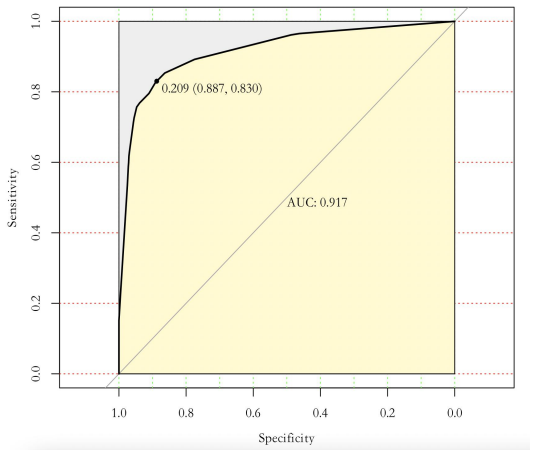

图 13 决策树模型预测的 ROC 曲线

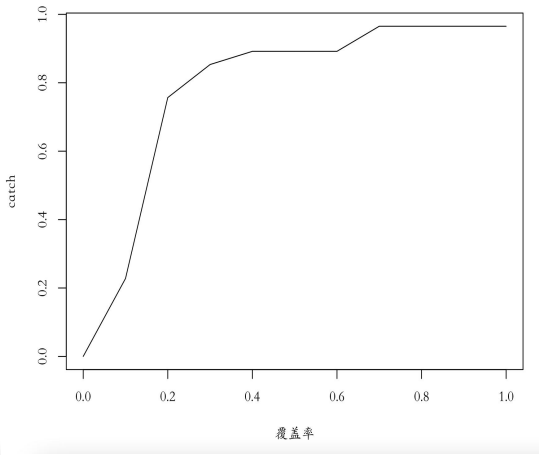

图 14 决策树模型预测的覆盖率-捕获率曲线

2、混淆矩阵

我们利用最佳阈值为 0.209 得到混淆矩阵及相关结果，如表 7:

表 7 最佳阈值的混淆矩阵

		预测值		
		非高价值客户	高价值客户	总计
真实值	非高价值客户	923	44	967
	高价值客户	117	215	332
	总计	1040	259	1299

由表 7 可以看出整体的错判率为 (44+117)/1299=12.4%，TPR 为 215/332=64.8%，FPR 为 44/967=4.55%。

3、准确率

通过建立决策树模型，我们最终得到模型准确率为 87.61%，说明模型的预测能力较强，可以认为模型性能良好。

（三）与 Logistic 回归模型的比较

综上所述，我们一共建立了 2 个模型，对比其评估指标，我们可以得到决策树模型的准确率略高于 Logistic 回归模型且 ROC 曲线更靠近左上方，混淆矩阵也展现的更好，因此可以判断决策树模型性能略优于 Logistic 模型，但两个模型都体现出了较好的性能，均可以作为筛选高价值客户的方法。

七、其他探索——基于 RFM 的高价值客户分类

前文我们已经简单介绍了 RFM 模型及其指标，除了帕累托法则给出的高价值客户定义，我们还可以通过 RFM 的打分体系，将客户群体分为 8 类，分别为重要价值客户、重要保持客户、重要发展客户、重要挽留客户、一般价值客户、一般保持客户、一般发展客户、一般挽留客户。其中，对于近度，我们设定高于

均值的记为低水平（0），低于均值的记为高水平（1），而频度和强度恰好相反，8 种客户类型具体对应的指标如表 8:

表 8　8 种客户类型及其 RFM 指标

R	F	M	RFM	客户类型
高	高	高	111	重要价值客户
高	高	低	110	一般价值客户
高	低	高	101	重要发展客户
高	低	低	100	一般发展客户
低	高	高	011	重点保持客户
低	高	低	010	一般保持客户
低	低	高	001	重点挽留客户
低	低	低	000	一般挽留客户

我们对数据进行 RFM 指标的构建，并进行 Kmeans 聚类，得到该电商平台的客户分布如下所示。可以看出客户群体中一般保持客户最多，而重要价值客户位列第四，这说明将普通客户转化为重要价值客户仍是一个有待解决的问题。而从聚类图中可以看出，我们可以大致将客户聚为 3 大类，聚类效果良好。其中第三类可以作为高价值客户。因此综上，我们也可以通过 RFM 客户评估体系再进行 Kmeans 的方法筛选高价值客户。

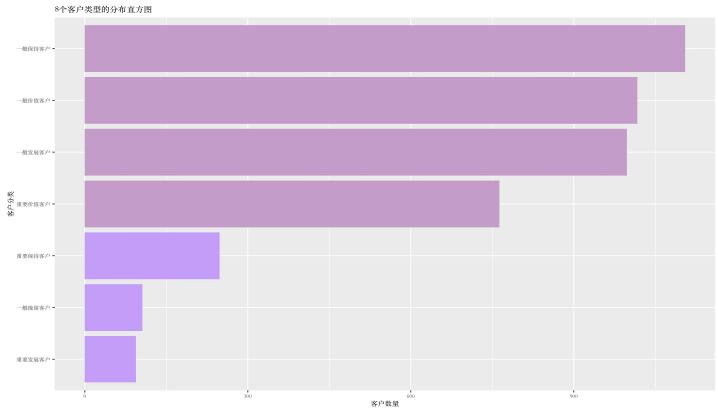

图 15　8 个客户类型的分布直方图

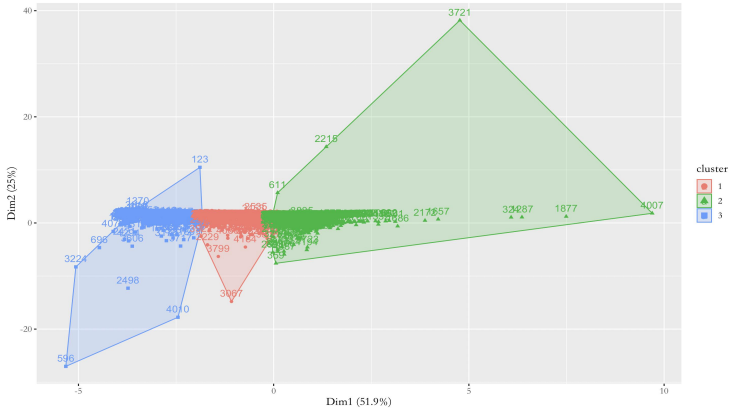

图 16　RFM 模型聚类可视化

八、结论与展望

本文针对商品价值评估和高价值客户筛选的问题，目标研究商品商业价值并找到高价值客户，对某电商平台订单的数据进行分析，利用 Kmeans 聚类评估商品的价值，并结合 RFM 指标建立 Logistic 回归模型和决策树模型，提出了一套筛选高价值客户的流程与方法。通过建模结果分析，本文一共得出如下结论与展望：

- 在评估商品价值时，基于 Kmeans 聚类，我们可以得到商品的畅销量、退货率、盈利能力都可以作为评估商品商业价值的指标，且可以很好的对商品进行聚类，本文我们将类别聚为 5 类，可以观察到每一类拥有的特征，以此来对每一类商品进行了解，很好的进行降维。
- 在筛选高价值客户时，我们想利用合适的指标来定位高价值客户，以此达到筛选目的。结合 RFM 三个指标以及购买量、消费金额、退单率这几个指标建立 Logistic 回归模型和决策树模型，得到分类正确率分别在 80.83%和 87.61%，AUC 值分别为 0.917，决策树模型略优于 Logistic 回归模型，两个模型拟合效果均良好，可以用于筛选高价值客户。
- 除了帕累托法则下的高价值客户，我们利用 RFM 指标对客户进行分类，可以采用 Kmeans 聚类的方式将高价值客户进行筛选，发现聚类效果也不错，因此，我们也可以通过 RFM 客户评估体系再进行 Kmeans 的方法筛选高价值客户。
- 借助 Logistic 回归模型和决策树模型，我们可以得出高价值客户的影响因素，可以通过 R、F、M、消费金额、退单率这几个指标来预测未来客户是否可能成为高价值客户。这可以使电商平台的商家优化自身的营销策略，吸取更多高价值客户，更好的促进电商平台的运营。