# 第12章 聚类分析

聚类分析是通过聚类方法将客户划分为不同的群体，对不同群体的客户进行特征分析，比较不同类别客户的价值，对不同价值的客户类别提供个性化服务，制定相应的营销策略。

## 12.1 聚类模型方法论

常用的聚类分析方法是基于RFM模型，通过K-Means聚类划分客群，然后从R、F、M三个维度分析客户的价值。客户价值分析的总体流程如图3.26所示。

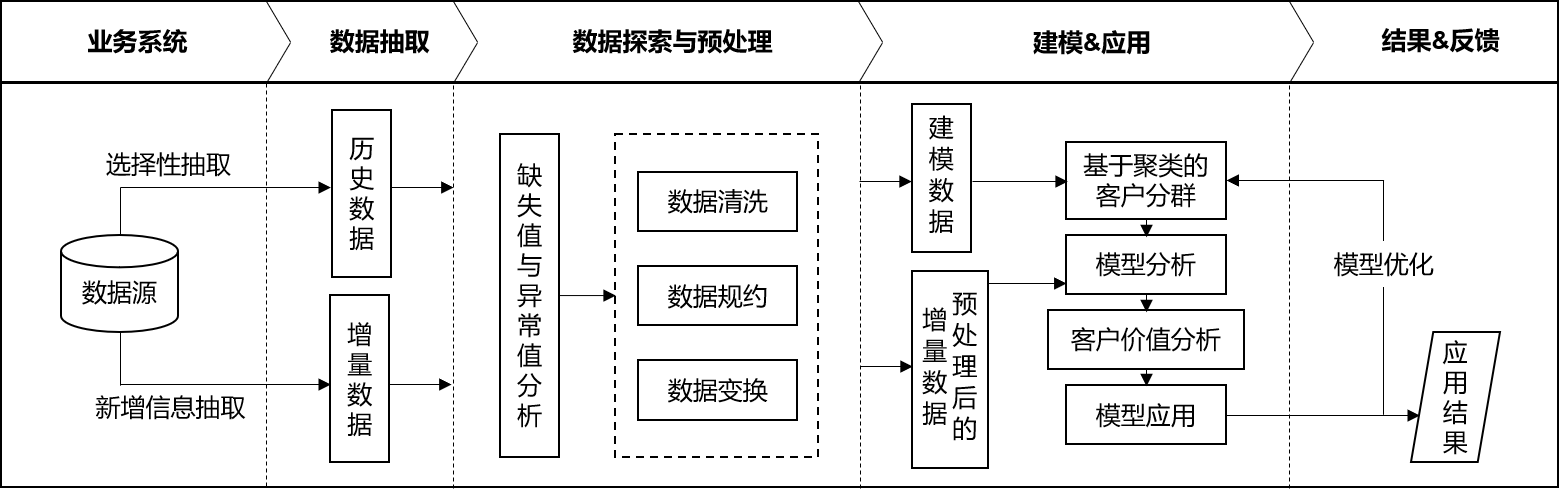


图3.26 客户价值分析的总体流程

注：RFM模型是衡量当前用户价值和客户潜在价值的重要工具和手段，R、F、M分别表示Rencency（最近一次消费），Frequency（消费频率）和Monetary（消费金额）[4]。

聚类分析的建模思路一般包括如下三个步骤：

1. **客户分类**：借助获取的客户数据，对客户进行分类；
2. **价值分析**：对不同的客户类别进行特征分析，比较不同类别客户的价值；
3. **制定策略**：对不同价值的客户类别提供个性化服务，制定相应的营销策略；

注：其中，客户分类方法不限于K-Means聚类，客户价值分析方法不限于RFM模型，评价维度(R、F、M)可根据实际应用领域定义。

## 12.2 K-Means聚类

K-Means算法是典型的基于距离的非层次聚类算法，在最小化误差函数的基础上将数据划分为设定的K个类别，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。

### 12.2.1 算法过程

1. 从N个样本数据中随机选取K个对象作为初始的聚类中心；
2. 分别计算每个样本到各个聚类中心的距离，将对象分配到距离最近的聚类中；
3. 所有对象分配完成后，重新计算K个聚类的中心；
4. 与前一次计算得到的K个聚类中心比较，如果聚类中心发生变化，转过程（2），否则转过程（5）；
5. 当质心不发生变化时停止并输出聚类结果；

聚类的结果可能依赖于初始聚类中心的随机选择，可能使得结果严重偏离全局最优分类。实践中，为了得到较好的结果，通常选择不同的初始聚类中心，多次运行K-Means算法。在所有对象分配完成后，重新计算K个聚类的中心时，对于连续数据，聚类中心取该簇的均值，但是当样本的某些属性是分类变量时，均值可能无定义，可以使用K-众数方法。

### 12.2.2 相似度度量

（1）连续属性

对于连续属性，要先对各属性值进行零-均值规范，再进行距离的计算。在K-Means聚类算法中，一般需要度量样本之间的距离、样本与簇之间的距离以及簇与簇之间的距离。

度量样本之间的相似性最常用的是欧几里得距离、曼哈顿距离和闵可夫斯基距离；样本与簇之间的距离可以用样本到簇中心的距离；簇与簇之间的距离可以用簇中心的距离。

用p个属性来表示n个样本的数据矩阵如下:



欧几里得距离：



曼哈顿距离：



闵可夫斯基距离：



q为正整数，q=1时即为曼哈顿距离；q=2时即为欧几里得距离。

（2）文档数据

对于文档数据使用余弦相似性度量，先将文档数据整理成文档-词矩阵格式，如表3-5所示。

表3.5 文档-词矩阵

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 词  文档 | lost | win | team | score | music | happy | sad | … | coach |
| 文档一 | 14 | 2 | 8 | 0 | 8 | 7 | 10 | … | 6 |
| 文档二 | 1 | 13 | 3 | 4 | 1 | 16 | 4 | … | 7 |
| 文档三 | 9 | 6 | 7 | 7 | 3 | 14 | 8 | … | 5 |

两个文档之间的相似度的计算公式为：



### 12.2.3 目标函数

使用误差平方和SSE作为度量聚类质量的目标函数，对于两种不同的聚类结果，选择误差平方和较小的分类结果。

连续属性的SSE计算公式为：



文档数据的SSE计算公式为：



簇E的聚类中心e计算公式为：



其中：

|  |  |  |  |
| --- | --- | --- | --- |
| 符号 | 含义 | 符号 | 含义 |
| K | 聚类簇的个数 |  | 簇的聚类中心 |
|  | 第i个簇 |  | 第i个簇中样本的个数 |
| x | 对象（样本） |  |  |

### 12.2.4 选择K值

选取K值时，有时候我们会根据行业经验或者需要来设定一个大概的聚类数，但是这不一定是我们获取到数据的真实聚类数。如果我们希望能从数据自身出发去确定真实的聚类数，也就是对数据而言的最佳聚类数，可以采用两种常用的确定聚类数k的方法。

（1）手肘法

一般采用手肘法选取最佳聚类数k，手肘法的核心思想是：随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和SSE自然会逐渐变小。并且，当k小于真实聚类数时，由于k的增大会大幅增加每个簇的聚合程度，故SSE的下降幅度会很大，而当k到达真实聚类数时，再增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减，然后随着k值的继续增大而趋于平缓，也就是说SSE和k的关系图是一个手肘的形状，而这个肘部对应的k值就是数据的真实聚类数。当然，这也是该方法被称为手肘法的原因。

具体做法是：让k从1开始取值直到取到你认为合适的上限(一般来说这个上限不会太大，这里我们选取上限为8)，对每一个k值进行聚类并且记下对应的SSE，然后画出k和SSE的关系图（毫无疑问是手肘形），最后选取肘部对应的k作为我们的最佳聚类数。

比如，对于某个数据集，k与SSE的关系图如图3.27所示：

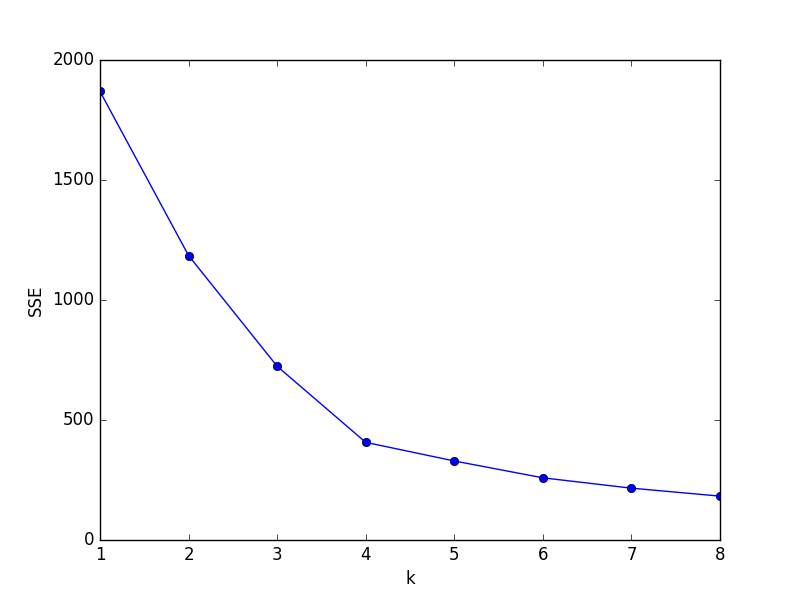


图3.27 k与SSE的关系

显然，肘部对应的k值为4，故对于这个数据集的聚类而言，最佳聚类数应该选4。

（2）轮廓系数法

该方法的核心指标是轮廓系数（Silhouette Coefficient），某个样本点Xi的轮廓系数定义如下：



其中，a是Xi与同簇的其他样本的平均距离，称为凝聚度，b是Xi与最近簇中所有样本的平均距离，称为分离度。而最近簇的定义是：

其中p是某个簇Ck中的样本。事实上，简单点讲，就是用Xi到某个簇所有样本平均距离作为衡量该点到该簇的距离后，选择离Xi最近的一个簇作为最近簇。

求出所有样本的轮廓系数后再求平均值就得到了平均轮廓系数。平均轮廓系数的取值范围为[-1,1]，且簇内样本的距离越近，簇间样本距离越远，平均轮廓系数越大，聚类效果越好。那么，很自然地，平均轮廓系数最大的k便是最佳聚类数。

使用同样的数据集，同样考虑k等于1到8的情况，对于每个k值进行聚类并且求出相应的轮廓系数，然后做出k和轮廓系数的关系图，选取轮廓系数取值最大的k作为我们最佳聚类系数。

聚类数k与轮廓系数的关系如图3.28所示：

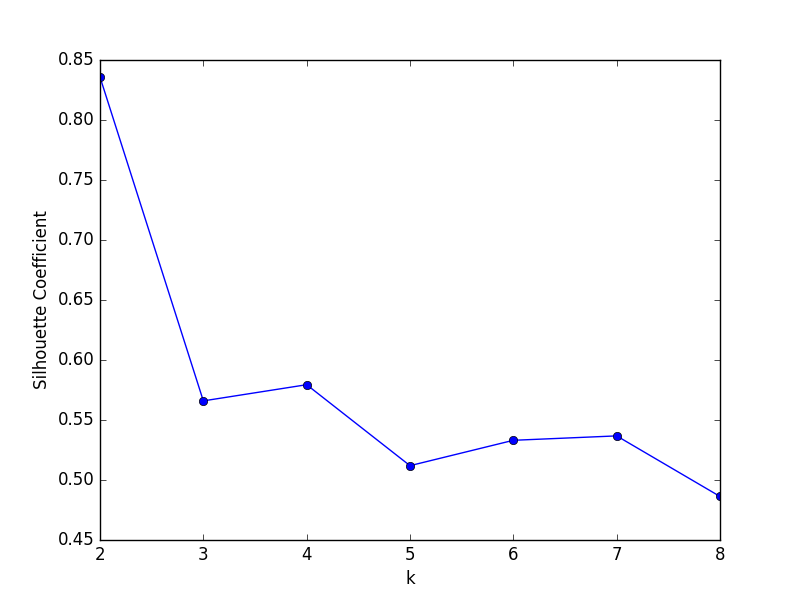


图3.28 聚类数k与轮廓系数的关系

可以看到，轮廓系数最大的k值是2，这表示我们的最佳聚类数为2。但是，值得注意的是，从k和SSE的手肘图可以看出，当k取2时，SSE还非常大，所以这是一个不太合理的聚类数，我们退而求其次，考虑轮廓系数第二大的k值4，这时候SSE已经处于一个较低的水平，因此最佳聚类系数应该取4而不是2。

但是，讲道理，k=2时轮廓系数最大，聚类效果应该非常好，那为什么SSE会这么大呢？在我看来，原因在于轮廓系数考虑了分离度b，也就是样本与最近簇中所有样本的平均距离。为什么这么说，因为从定义上看，轮廓系数大，不一定是凝聚度a（样本与同簇的其他样本的平均距离）小，而可能是b和a都很大的情况下b相对a大得多，这么一来，a是有可能取得比较大的。a一大，样本与同簇的其他样本的平均距离就大，簇的紧凑程度就弱，那么簇内样本离质心的距离也大，从而导致SSE较大。所以，虽然轮廓系数引入了分离度b而限制了聚类划分的程度，但是同样会引来最优结果的SSE比较大的问题，这一点也是值得注意的。

## 12.3 t-SNE可视化算法

t-SNE是用于降维的一种机器学习算法，是一种非线性降维算法，非常适用于高维数据降维到2维或者3维，进行可视化。并且t-SNE主要是关注数据的局部结构[6]。

总的来说，t-SNE算法的原理就是将数据点之间的相似度转换为概率，并且保证原始数据与降维后数据的概率分布尽可能保持一致，既达到降维的目的，又能使不同类别的样本点分开，如图3.29所示。通过原始空间和嵌入空间的联合概率的 Kullback-Leibler（KL）散度来评估可视化效果的好坏，也就是说用有关 KL 散度的函数作为 loss 函数，然后通过梯度下降最小化 loss 函数，最终获得收敛结果。并且t-SNE在低维空间下使用更重长尾分布的t分布来避免crowding问题和优化问题[5]。

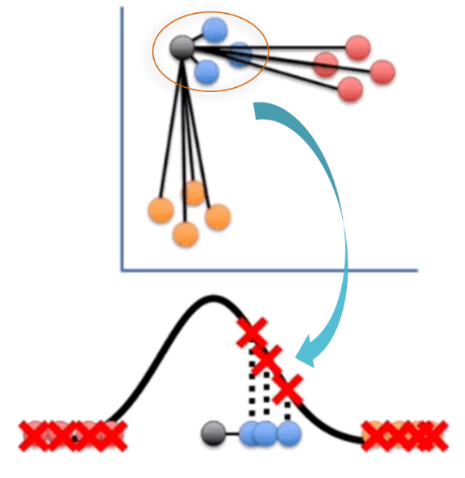


图3.29 t-SNE算法的相似度转换

注：t-SNE 的计算复杂度很高，在数百万个样本数据集中可能需要几个小时，而 PCA 可以在几秒钟或几分钟内完成。

## 12.4 赋予类别业务含义

通过特征分析，对不同类别比较各特征的分布图或雷达图，找出优势特征，为每个类别赋予实际的业务含义。如图3.30所示，可以比较不同客群的优势特征，为每个客群赋予实际含义。

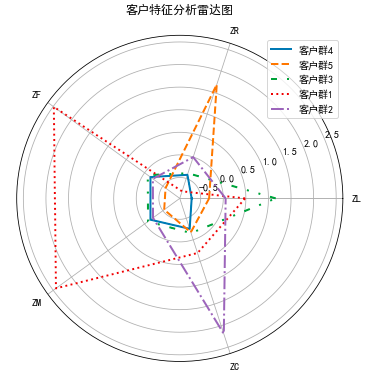
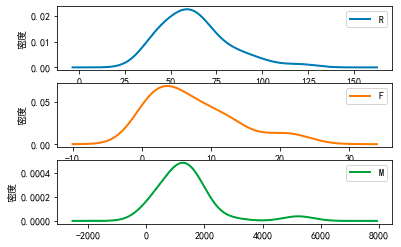


图3.30 不同客群的特征分析

## 12.5 项目演示：航空客户价值分析

（1）业务背景

本案例的目标是客户价值识别，即通过航空公司客户数据识别不同价值的客户[2]。识别客户价值应用最广泛的模型是通过3个指标（最近消费时间间隔（Recency）、消费频率（Frequency）和消费金额（Monetary））来进行客户细分，识别出高价值的客户，简称RFM模型。

在RFM模型中，消费金额表示在一段时间内，客户购买该企业产品金额的总和。由于航空票价受到运输距离、舱位等级等多种因素影响，同样消费金额的不同旅客对航空公司的价值是不同的。例如，一位购买长航线、低等级舱位票的旅客与一位购买短航线、高等级舱位票的旅客相比，后者对于航空公司而言价值可能更高。因此，这个指标并不适用于航空公司的客户价值分析。我们选择客户在一定时间内累积的飞行里程M和客户在一定时间内乘坐舱位所对应的折扣系数的平均值C两个指标代替消费金额。此外，考虑航空公司会员入会时间的长短在一定程度上能够影响客户价值，所以在模型中增加客户关系长度L，作为区分客户的另一指标。

本案例将客户关系长度L、消费时间间隔R、消费频率F、飞行里程M和折扣系数的平均值C五个指标作为航空公司识别客户价值指标（见表3-6），记为LRFMC模型。

表3-6 航空公司衡量客户价值的维度

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型 | L | R | F | M | C |
| 航空公司LRFMC模型 | 会员入会时间距观察窗口结束的月数 | 客户最近一次乘坐公司飞机距观测窗口结束的月数 | 客户在观测窗口内乘坐公司飞机的次数 | 客户在观测窗口内累计的飞行里程 | 客户在观测窗口内乘坐舱位所对应的折扣系数的平均值 |

针对航空公司LRFMC模型，如果采用传统RFM模型分析的属性分箱方法，虽然也能够识别出最有价值的客户，但是细分的客户群太多，提高了针对性营销的成本。因此，本案例采用聚类的方法识别客户价值。通过对航空公司客户价值的LRFMC模型的五个指标进行K-Means聚类，识别出最有价值客户。

该案例的客户价值分析模型主要由两个部分构成，第一个部分根据航空公司客户5个指标的数据，对客户进行聚类分群。第二部分结合业务对每个客户群进行特征分析，分析其客户价值，并对每个客户群进行排名。

（2）客户分群

采用K-Means聚类算法对客户数据进行客户分群，聚成5类（需要结合业务的理解与分析来确定客户的类别数量）。聚类后分群的结果如表3-7所示：

表3-7 不同客群的聚类中心

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 聚类类别 | 聚类个数 | 聚类中心 | | | | |
| ZL | ZR | ZF | ZM | ZC |
| 客户群1 | 5337 | 0.483 | -0.799 | 2.483 | 2.424 | 0.308 |
| 客户群2 | 15735 | 1.160 | -0.377 | -0.087 | -0.095 | -0.158 |
| 客户群3 | 12130 | -0.314 | 1.686 | -0.574 | -0.537 | -0.171 |
| 客户群4 | 24644 | -0.701 | -0.415 | -0.161 | -0.165 | -0.255 |
| 客户群5 | 4198 | 0.057 | -0.006 | -0.227 | -0.230 | 2.191 |

通过t-SNE算法将聚类后的数据进行客户化，得到如图3.31所示的结果。从图中可以看出，不同类别的数据被分开，并且类间的交互性不高，聚类结果具有实际价值，可以进行下一步群体分析。

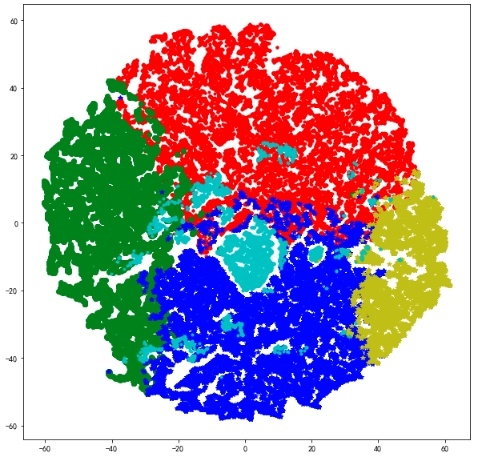


图3.31 聚类结果的可视化展示

（3）群体分析/客户价值分析

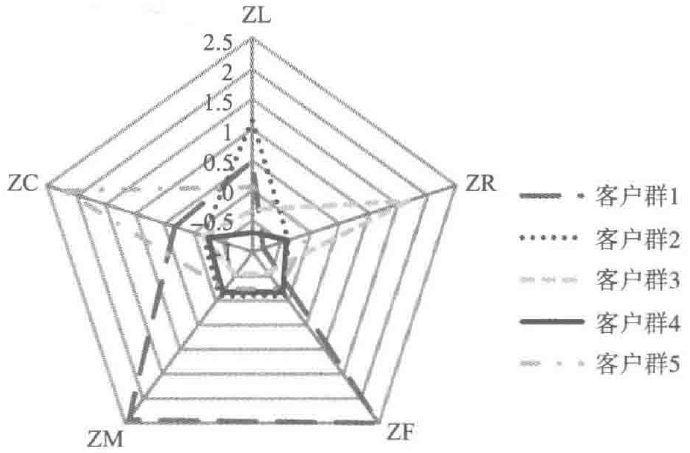


图3.32 客户群特征分析

针对聚类结果进行特征分析，如图所示。其中，客户群1在F、M属性上最大，在R属性上最小；客户群2在L属性上最ZL大；客户群3在R属性上最大，在F、M属性上最小；客户群4在L、C属性上最小；客户群5在C属性上最大。结合业务分析，通过比较各个指标在群间的大小对某一个群的特征进行评价分析。例如客户群1在F、M属性最大，在R指标最小，因此可以说F、M、R在客户群1是优势特征。以此类推，F、M、R在客户群3上是劣势特征。从而总结出每个群的优势和弱势特征，具体结果如表3-8所示。

表3-8 客户群特征描述表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 群类别 | 优势特征 | | | 弱势特征 | | | |
| 客户群1 | F | M | R | - | | | |
| 客户群2 | L | F | M | - | | | |
| 客户群3 | - | | | F | M | | R |
| 客户群4 | - | | | L | | C | |
| 客户群5 | C | | | R | F | | M |

由上述的特征分析的图表说明每个客户群的都有显著不同的表现特征，基于该特征描述，本案例定义五个等级的客户类别：重要保持客户、重要发展客户、重要挽留客户、一般客户、低价值客户。每种客户类别的特征如下：

* **重要保持客户**：这类客户的平均折扣率（C）较高（一般所乘航班的舱位等级较高），最近乘坐过本公司航班（R）低，乘坐的次数（F）或里程（M）较高。他们是航空公司的高价值客户，是最为理想的客户类型，对航空公司的贡献最大，所占比例却较小。航空公司应该优先将资源投放到他们身上，对他们进行差异化管理和一对一营销，提高这类客户的忠诚度与满意度，尽可能延长这类客户的高水平消费。
* **重要发展客户**：这类客户的平均折扣率（C）较高，最近乘坐过本公司航班（R）低，但乘坐次数（F）或乘坐里程（M）较低。这类客户入会时长（L）短，他们是航空公司的潜在价值客户。虽然这类客户的当前价值并不是很高，但却有很大的发展潜力。航空公司要努力促使这类客户增加在本公司的乘机消费和合作伙伴处的消费，也就是增加客户的钱包份额。通过客户价值的提升，加强这类客户的满意度，提高他们转向竞争对手的转移成本，使他们逐渐成为公司的忠诚客户。
* **重要挽留客户**：这类客户过去所乘航班的平均折扣率（C）、乘坐次数（F）或者里程（M）较高，但是较长时间已经没有乘坐本公司的航班（R）高或是乘坐频率变小。他们客户价值变化的不确定性很高。由于这些客户衰退的原因各不相同，所以掌握客户的最新信息、维持与客户的互动就显得尤为重要。航空公司应该根据这些客户的最近消费时间、消费次数的变化情况，推测客户消费的异动状况，并列出客户名单，对其重点联系，采取一定的营销手段，延长客户的生命周期。
* **一般与低价值客户**：这类客户所乘航班的平均折扣率（C）很低，较长时间没有乘坐过本公司航班（R）高，乘坐的次数（F）或里程（M）较低，入会时长（L）短。他们是航空公司的一般用户与低价值客户，可能是在航空公司机票打折促销时，才会乘坐本公司航班。

其中，重要发展客户、重要保持客户、重要挽留客户这三类重要客户分别可以归入客户生命周期管理的发展期、稳定期、衰退期三个阶段。

根据每种客户类型的特征，对各类客户群进行客户价值排名，其结果如表3-9所示。针对不同类型的客户群提供不同的产品和服务，提升重要发展客户的价值、稳定和延长重要保持客户的高水平消费、防范重要挽留客户的流失并积极进行关系恢复。

表3-9 客户群价值排名

|  |  |  |
| --- | --- | --- |
| 客户群 | 排名 | 排名含义 |
| 客户群1 | 1 | 重要保持客户 |
| 客户群5 | 2 | 重要发展客户 |
| 客户群2 | 3 | 重要挽留客户 |
| 客户群4 | 4 | 一般客户 |
| 客户群3 | 5 | 低价值客户 |