航空公司客户价值分析

报告人：熊欣

时间：2019 年 11 月 27 日

目录

1 问题描述

客户关系管理信息时代企业的核心问题。其关键任务是客户分类，通过分析客户价值，针对不同价值客户制定个性化服务方案与营销策略，将有效营销资源集中与高价值客户，实现企业利润最大化。国内某航空公司面临客户流失，竞争力下降和航空资源未充分利用等经营危机。该公司拟通过数据分析手段，建立合理的客户价值评估模型，对客户进行分群，针对不同群组下客户的价值，制定营销策略来解决目前公司目前面临的一系列问题。

本实验利用 K-means 算法，对提供的航空公司的客户数据进行客户聚类。该过程中包含采用内部评测法对聚类效果进行评价以选择合理的 K 值，对汇聚的客户群组有较好的解释。本次实验包括以下三个分析目标：

（1）借助航空公司客户数据，对客户进行 K-means 聚类。。

（2）对不同的客户群组的客户特征进行分析，比较不同类客户的客户价值。

（3）尝试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。

2 方法描述

本实验目标是通过航空公司的客户数据识别不同价值的客户。在客户价值分析中使用较多的 RFM 模型，提出发现客户数据中有三大重要指标：最近一次消费频率（Recency）、消费频率（Frequency）、消费金额（Monetary），从而识别出高价值的客户。

但在本案例中，RFM 模型不能完全适用。如消费金额表示在一段时间内的客户消费金额总和，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乘客对于航空公司的价值并不相同。即购买短航线、高等级舱位飞机票的旅客，相较于购买长航线、低等级舱位机票的旅客的客户价值要高。因此我们将该指标改进为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。此外考虑航空公司会员入会时间的长短在一定程度上也能影响客户价值，故增设客户入会的关系长度 L 这一指标。最终本案例使用的 LRFMC 模型中，作为航空公司识别客户价值的五个指标为：会员入会时间距观测窗口结束的月份 L、客户最近一次乘坐公司飞机距观测窗口结束的月份 R、客户在观测窗口内乘坐公司飞机的次

数 F、客户在观测窗口内累计飞行里程 M、客户在观测窗口内乘坐仓位所对应的折扣系数的均值 C。

传统 RFM 模型分析的属性分箱方法如下图 1 所示，通过观察 R、F、M 指标的值来识别最有价值的客户。但本案例的 LRFMC 模型从五个维度上划分客户群太多，故考虑采用 K-means 方法对这五个指标进行聚类，从而实现客户价值分析，分析的总体流程如图 2 所示。

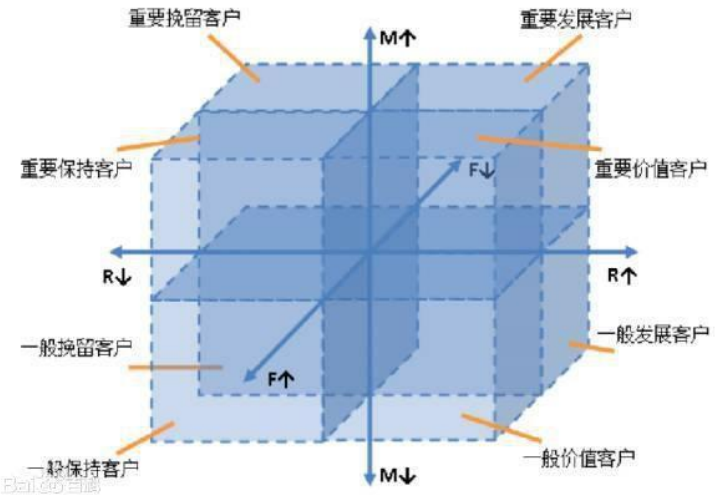

图 1 传统 RFM 模型分析

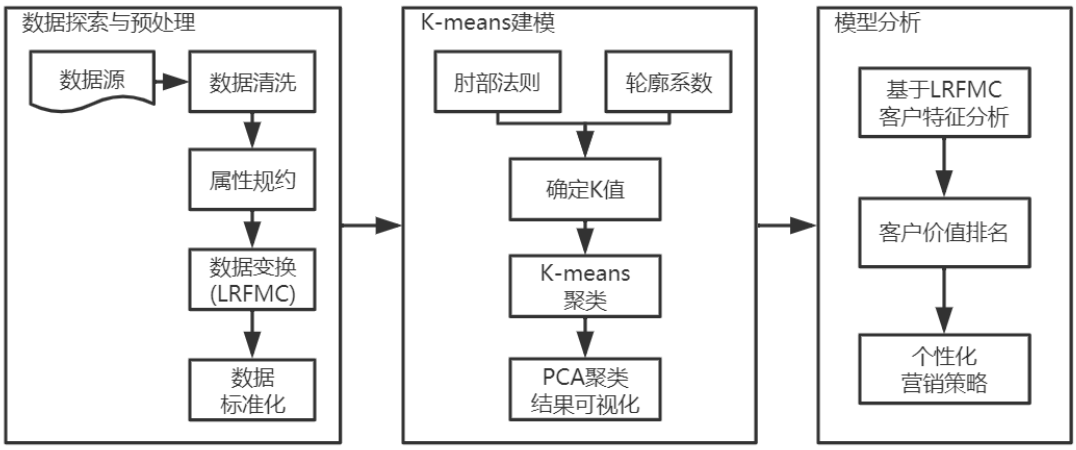

图 2 航空公司客户价值分析总体流程

航空公司客户价值分析实验过程主要分为三个部分，第一个部分是对数据 air_data.csv 的前期处理，包括从原始数据源中抽取待分析的历史观测数据，清洗含空值的数据记录，基于构建的 LRFMC 模型对数据进行属性规约和变换，最后对数据进行标准化处理，消除数量级数据带来的影响。

第二个部分是 K-means 建模，首先利用内部评测方法——肘部法则和轮廓系数对聚类效果进行评价，同时结合实际问题特点选择合理 K 值。然后选择 K=5 时进行 K-means 聚类，查看航空公司客户聚类分群结果表。同时为了可视化聚类结果，本实验利用 PCA 主成分分析法将特征维度降至两维呈现其聚类结果。

第三个部分是模型的应用和分析，包括对聚类的五类客户群进行基于 LRFMC 指标的客户特征分析，对每类客户群组进行有效解释并作客户价值排名。最后基于客户价值分析结果提出一些可参考性的个性化营销策略，提升航空公司客户的忠诚度和满意度。

3 数据分析与建模

本章节主要对原始数据进行探索性分析，及数据清洗、数据规约、数据标准化等预处理操作，然后采用内部评测法对聚类效果进行评价以选择合理 K 值，再进行 K-means 聚类模型的构建。

3.1 数据探索与预处理

3.1.1 数据清洗

在该实验数据集上总计 62988 条数据记录中，以末次飞行日期（LAST_FLIGHT_DATE）2014-03-31 为结束时间，选取宽度为两年的时间段即 2012-4-1 至 2014-3-31，作为观测窗口，抽取观测窗口内所有客户的详细数据，形成历史数据；对于后续新增的客户信息，以新增数据中最新的时间点作为结束时间，形成新增数据。

对本实验数据的 44 个属性进行探索分析，包含缺失值与异常值的分析，观察探索原始数据中存在空值或最小值为 0 的情况原因。其中票价收入为空值可能是客户不存在乘机记录造成，票价收入最小值为 0、折扣率最小值为 0 而总飞行公里数大于 0 可能是客户乘坐积分兑换或 0 折机票产生的。由于原始数据量较大，脏数据所占比例较小，故删除票价收入（SUM_YR）为空，票价收入（SUM_YR）为 0、平均折扣率（AVG_DISCOUNT）不为 0 且总飞行公里数（SEG_KM_SUM）大于 0 的逻辑异常数据记录，得到最终 62032 条数据记录。

3.1.2 数据规约

由于原始数据集中属性列太多，根据第二章构建的航空公司客户价值分析 LRFMC 模型，选择与 LRFMC 指标相关的 6 个属性：FFP_DATE（入会时间）、LOAD_TIME（观测窗口的结束时间）、FLIGHT_COUNT（观测窗口内的飞行次数）、AVG_DISCOUNT（平均折扣率）、SEG_KM_SUM（观测窗口的总飞行公里数）、LAST_TO_END（最后一次乘机至观测窗口结束时长）。删除不相关或弱相关冗余的属性，如 MEMBER_NO（会员卡号）、GENDER（性别）、WORK_CITY（工作地城市）、AGE(年龄)等属性。

3.1.2 数据标准化

由于原始数据中并未直接提供 LRFMC 模型中的五个指标，需要通过数据变换操作从原始数据中计算提取出这五个指标。

L（会员入会时间距观测窗口结束的月份）= LOAD_TIME（观测窗口的结束时间）-FFP_DATE（入会时间）

R（客户最近一次乘坐公司飞机距观测窗口结束的月份）= LAST_TO_END（最后一次乘机至观测窗口结束时长）

F（客户在观测窗口内乘坐公司飞机的次数）= FLIGHT_COUNT（观测窗口内的飞行次数）

M（客户在观测窗口内累计飞行里程）= SEG_KM_SUM（观测窗口的总飞行公里数）

C（客户在观测窗口内乘坐仓位所对应的折扣系数的均值）= AVG_DISCOUNT（平均折扣率）

由于这五个指标的数据分布取值范围差异较大，为了消除数量级数据带来的影响，需要对数据进行标准化处理。

3.2 K-means 聚类模型构建

本实验中先对航空公司客户价值 LRFMC 模型的五个指标进行聚类分群分析，再对聚类后的每个客户群进行特征分析其客户价值。利用 K-means 聚类算法对客户数据进行分群聚类，需要事先确定聚类数目 K 值。关于如何确定本实验中聚类模型的合理 K 值，主要使用到两种内部评测方法对聚类效果进行评

价，即肘部法则和轮廓系数评估计算。同时结合本案例的实际需求，为航空公司的客户数据进行分群聚类并进行客户价值分析，来确定最终合理 K 值。

肘部法则的计算原理是成本函数，成本函数是类别畸变程度之和，每个类的畸变程度等于每个变量点到其类别中心的位置距离平方和，若类内部的成员彼此间越紧凑则类的畸变程度越小，反之亦然。在选择类别数量上，肘部法则会把不同值的成本函数值画出来。随着值的增大，平均畸变程度会减小；每个类包含的样本数会减少，于是样本离其重心会更近。但是，随着值继续增大，平均畸变程度的改善效果会不断减低。值增大过程中，畸变程度的改善效果下降幅度最大的位置对应的值就是肘部。本文设置簇类数取值区间为 1~10，计算各结点到类中心距离的均值绘制折线图如下图 3 所示，可以发现随着 K 值的增加，各结点到类中心距离的均值逐渐减小，趋于稳定，下降幅度最大的位置即肘部对应的 K 值为 4。

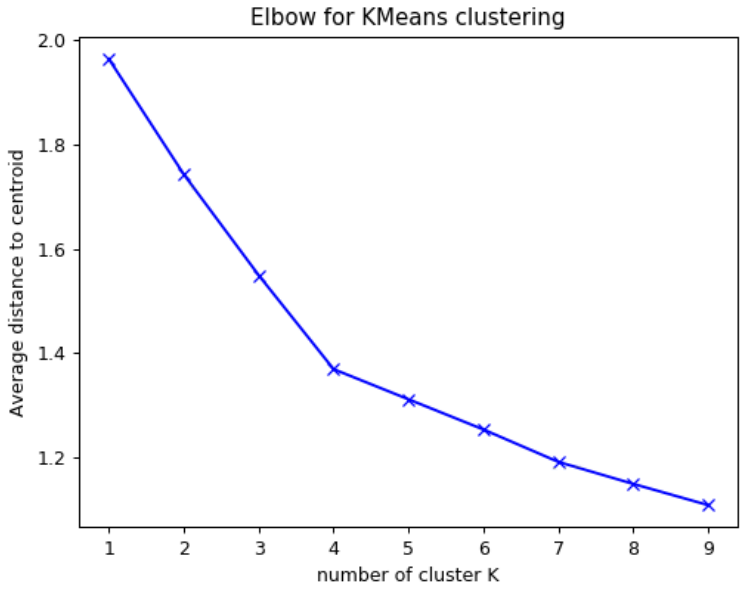

图 3 K-means 聚类的肘部法则图

轮廓系数是度量类的密集与分散程度，它会随着类的规模增大而增大。彼此相距很远，本身很密集的类，其轮廓系数较大，反之亦然。则轮廓系数越大，表明类与类间的离散型越好即聚类效果越好。本文设置簇类数取值区间为 2~10，计算类与类之间的轮廓系数值绘制折线图如下图 4 所示。由于 K-means 聚类是随机选择重心点，即随机从不同位置开始初始化，故一般形成局部最优

的聚类结果。结合实际应用中对客户分群 2 类进行价值分析的意义不大，故本实验中利用轮廓系数确定的 K-means 聚类局部最优 K 值为 6。

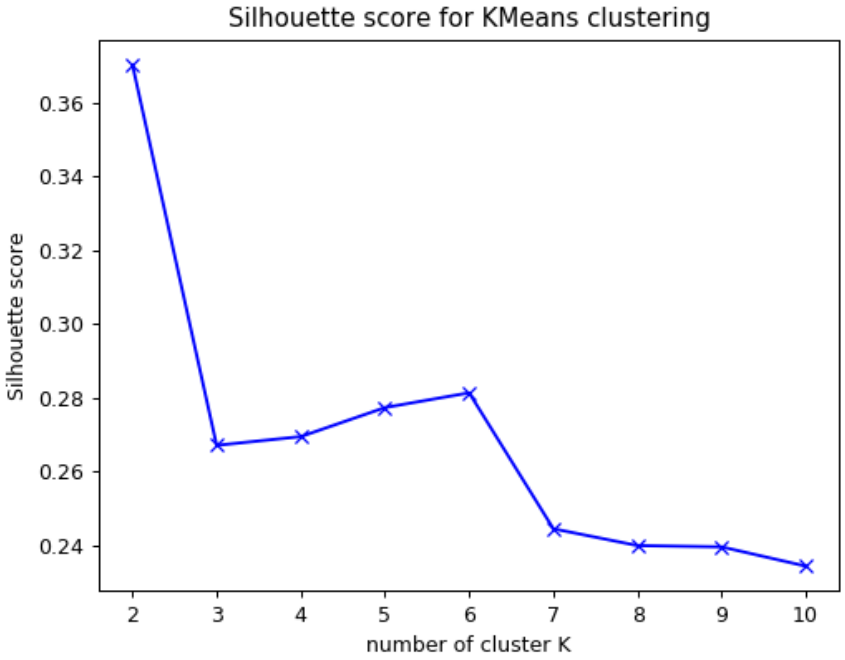

图 4 K-means 聚类的轮廓系数图

结合这两种评测方法选取的最优 K 值，考虑本案例的航空公司客户价值分析任务，本实验最终取肘部法则和轮廓系数选取的 K 值平均值，即 K=5 时，两种评测方法的综合效果最优。则输入清洗后的数据集，聚类类别数为 5，进行 K-means 聚类算法的代码如下。计算五个客户群类别的聚类个数以及聚类中心值，并将其聚类分群的结果展现如下表 1 所示。

```
# K-means 聚类算法(K=5)
import pandas as pd
from sklearn.cluster import KMeans
model = KMeans(n_clusters=5)
model.fit(cleanData)
# 查看聚类中心和各类别的聚类个数
r1 = pd.Series(model.labels_).value_counts()
r2 = pd.DataFrame(model.cluster_centers_)
r = pd.concat([r2, r1], axis=1)
r.columns = list(data4.columns) + ['Category Amount']
```

表 1 航空公司客户聚类分群结果表

聚类类别	聚类中心					聚类个数
	SL	SR	SF	SM	SC	
客户群 1	0.050641	-0.002818	-0.227136	-0.231885	2.189506	4188
客户群 2	-0.700182	-0.414936	-0.161121	-0.160899	-0.255283	24651
客户群 3	1.160831	-0.377150	-0.086930	-0.094839	-0.156149	15735
客户群 4	0.483446	-0.799400	2.482949	2.424488	0.308702	5336
客户群 5	-0.313717	1.686278	-0.574023	-0.536819	-0.173296	12122

在对聚类结果的可视化上，由于用原始数据构建的 LRFMC 模型包含五维特征，维度太高不方便进行可视化操作。故一般使用 PCA 主成分分析方法来降低数据维度到 2 维平面，加快模型的学习收敛速度同时实现聚类结果的可视化展示，如下图 5 所示。可以发现图中呈现红、黄、蓝、绿、青五种颜色的分布，其中红、绿、蓝、黄色的聚类分布区别明显，青色存在较大面积的重叠分布，整体来看该模型的 K-means 聚类效果较好。

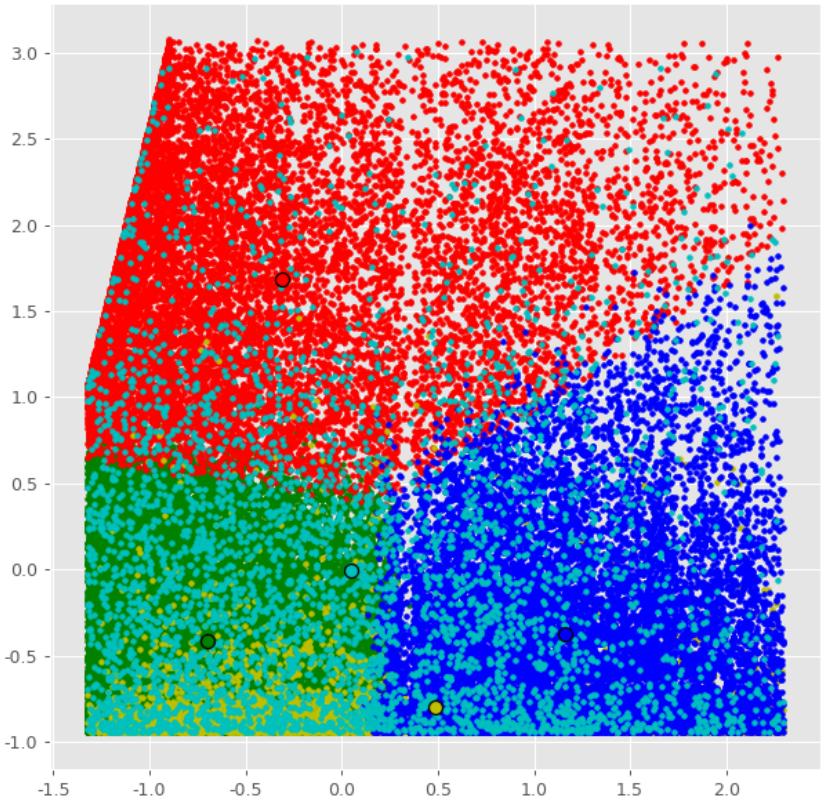

图 5 基于 PCA 的航空公司客户聚类结果可视化

4 实验结果讨论

4.1 客户价值分析

针对聚类的五个客户群类别进行基于 LRFMC 模型的特征分析，其中五个指标为 L 会员入会时间距观测窗口结束的月份；R 客户最近一次乘坐公司飞机距观测窗口结束的月份；F 客户在观测窗口内乘坐公司飞机的次数；M 客户在观测窗口内累计飞行历程；C 客户在观测窗口内乘坐仓位所对应的折扣系数的均值。可以从下图 6 中发现：

客户群 1（Customers 1）在 C 指标上最大，R 指标较大，其余指标一般，可定义为挽留客户；

客户群 2（Customers 2）在 L、C 指标上最小，其余指标一般，可定义为一般客户；

客户群 3（Customers 3）在 L 指标上最大，其余指标一般，可定义为普通保持客户；

客户群 4（Customers 4）在 F、M 指标上最大，L、C 指标较大，在 R 指标上最小，可定义为重要保持客户；

客户群 5（Customers 5）在 R 指标上最大，在 F、M 指标上最小，C、L 指标较小，可定义为低价值客户。

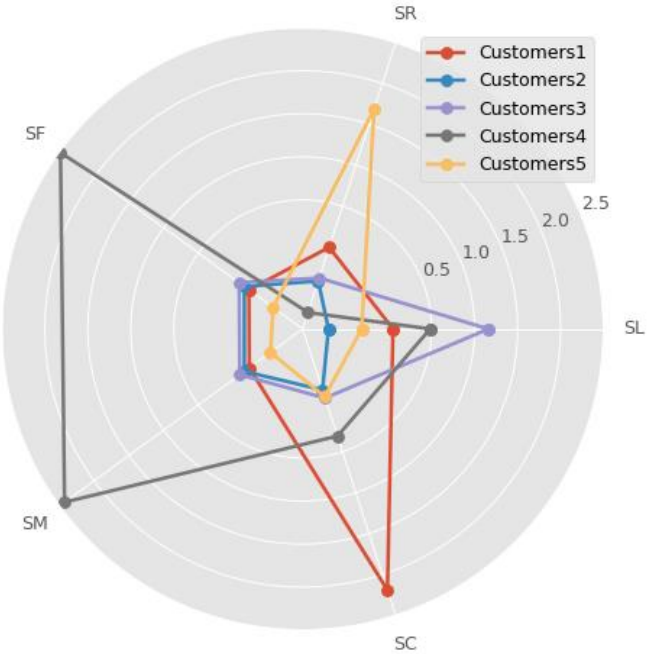

图 6 客户群特征分析图

由上述五类客户群展现的不同特征指标分析，本案例将对五个类别的客户群进行类型特征定义，分别为重要保持客户，重要发展客户，重要挽留客户，一般客户和低价值客户。其中对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（即航班舱位等级较高），乘坐次数 F、飞行里程 M 较高，最近一次乘坐航班 R 较低。他们属于航空公司的高价值客户，且所占比例较小。航空公司应将资源优先投放到这类客户身上，进行差异化管理，提高客户的忠诚度和满意度，使得这类客户继续保持高水平消费状态。

普通保持客户：这类客户的入会时长 L 长，但乘坐次数 F 或飞行里程 M 一般，平均折扣率 C 和最近一次乘坐航班 R 较低。这类客户整体贡献价值尚可且入会时间长，属于需要保持的普通航空公司客户。航空公司可以考虑通过客户价值的提升来加强该类客户的满意度，增加其在本公司的机票消费和合作伙伴处的消费，使他们逐渐成为公司的忠诚客户。

挽留客户：这类客户的平均折扣率 C、乘坐次数 F 或飞行里程 M 较高，最近一次乘坐航班 R 较高（即较长时间未乘坐该航空公司的航班）。该类客户价值变化的不确定性很高，航空公司应掌握客户最新信息、维持与客户的互动，了解推测客户乘坐航班减少的原因，采取一定营销手段增强与客户之间的联系，从而延长客户的生命周期。

一般和低价值客户：这类客户的平均折扣率 C、乘坐次数 F、飞行里程 M 较低，最近一次乘坐航班 R 较高（即较长时间未乘坐该航空公司的航班），入会时长 L 较短。航空公司可能在实行机票打折促销活动时，会吸引该类一般或低价值客户的消费。

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如下表 X 所示。后续针对不同类型的客户群提供不同的产品和服务，从而提升整体的客户价值。由于本模型是采用历史观测数据进行建模，随着时间变化需要对观测数据进行更新，来重新训练模型并进行调整。

表 X 客户价值排名

客户群	客户价值排名	排名含义
客户群 4	1	重要保持客户
客户群 3	2	普通保持客户

客户群 1	3	挽留客户
客户群 2	4	一般客户
客户群 5	5	低价值客户

4.2 个性化营销策略

根据对五个客户群的特征分析结果，提供以下个性化营销策略，为航空公司的客户群价值分析管理作参考。

（1）差异化营销。目前航空市场中，消费者对于价格和舒适性的需求之间的差异是非常明显的，部分乘客对于价格较为敏感，部分乘客对于乘坐的环境较为看重，还有部分乘客对于价格和环境舒适性都较敏感。因此航空公司应该充分结合这些差异性因素，指定差异化营销策略。针对重要保持和普通保持客户，航空公司应注重该类客户的服务质量水平。航空公司可以在飞机上加强多媒体设施的安装，通过设立移动媒体，建立手机连接来完善消费者对于信息的获取渠道。还可以在飞机上提供一些特色服务，例如：对于天气温度的实时播报、以及当地旅游景点的推荐，可以的帮助航空公司树立自己的品牌形象。针对挽留客户可以发放福利兑换或折扣优惠等活动，吸引他们的注意从而转化为普通保持客户。而针对一般或低价值客户，不需要投放太多资源在这类客户上，只需保持正常的机票促销活动投放。

（2）划分会员等级。航空公司可以为不同类型的会员客户划分不同等级，如金卡、银卡、普通卡会员等等。一般来说，会员制的管理方式基本上是在一定时间内积累的飞行里程数，数值高则能晋升成为更高级会员，并享受高级别服务。但许多客户对会员升级和保级的时间点和要求并不完全了解，其相关文件说明常常由于过于复杂不易理解，故经常会出现错过升级或保级的机会。因此，航空公司可以在会员升级或保级的结算时间节点，为相关客户推送提醒或促销活动，促进客户的会员升级或保级积极性和有效性，从而提升客户忠诚度和满意度。

（3）企业合作销售。航空公司可以考虑与非航空类的企业进行合作，使客户在航班过程中可以了解其他非航空类企业产品服务信息，同时在其他企业消费过程中获得航空公司会员积分等福利，增强客户与航空公司的联系紧密度。通过此企业间的合作销售模式，还可以了解重要客户在合作的非航空类公司中

的消费习惯和记录情况，从中挖掘出客户的消费行为特点，为他们提供相应的个性化促销活动。对于所乘航班的平均折扣率 C 较高，乘坐次数 F 较高，飞行里程 M 较高，最近一次乘坐航班 R 较低的高价值客户，航空公司应注重该类贡献量大的客户群体，优先将资源投放到他们身上，对他们进行有效地个性化营销管理，尽可能保持该类客户的高水平消费状态，提升他们的忠诚度和满意度。

5 实验总结

本实验通过问题描述对航空公司的会员数据进行探索和预处理操作，然后利用肘部法则和轮廓系数的内部评测方法来确定合适的 K 值，构建基于 LRFMC 客户价值分析模型的 K-means 聚类模型。并利用 PCA 主成分分析法对聚类结果进行可视化呈现，最后针对客户群的特征分析和价值排序，提出一些可参考性的个性化营销策略。由于获取新客户的成本远高于维持老客户，故航空公司应加强对高价值老客户的关系维系，进行差异化个性管理，延长该类客户高水平消费状态，整体提高客户满意度和忠诚度。

后续实验可以从以下几个方面进一步深入分析：尝试多种 K 值聚类结果方案，并深层次的分析客户群体特征，针对每一类客户群进行特定的营销方案设计。

附：程序代码

见文件“航空公司客户价值分析-熊欣.py”。