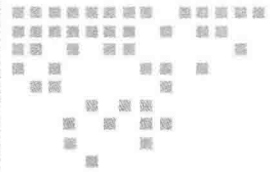

Chapter 7 第 7 章

航空公司客户价值分析

7.1 背景与挖掘目标

信息时代的来临使得企业营销焦点从产品中心转变为客户中心，客户关系管理成为企业的核心问题。客户关系管理的关键问题是客户分类，通过客户分类，区分无价值客户、高价值客户，企业针对不同价值的客户制定优化的个性化服务方案，采取不同营销策略，将有限营销资源集中于高价值客户，实现企业利润最大化目标。准确的客户分类结果是企业优化营销资源分配的重要依据，客户分类越来越成为客户关系管理中亟待解决的关键问题之一。

面对激烈的市场竞争，各个航空公司都推出了更优惠的营销方式来吸引更多的客户，国内某航空公司面临着旅客流失、竞争力下降和航空资源未充分利用等经营危机。通过建立合理的客户价值评估模型，对客户进行分群，分析比较不同客户群的客户价值，并制定相应的营销策略，对不同的客户群提供个性化的客户服务是必须和有效的。目前该航空公司已积累了大量的会员档案信息和其乘坐航班记录，经加工后得到表 7-1 所示的部分数据信息。

请根据这些数据（见表 7-2）实现以下目标。

1）借助航空公司客户数据，对客户进行分类。

2）对不同的客户类别进行特征分析，比较不同类客户的客户价值。

3）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。

表7-1 航空信息属性表

	属性名称	属性说明
客户基本信息	MEMBER_NO	会员卡号
	FFP_DATE	入会时间

（续）

	属性名称	属性说明
客户基本信息	FIRST_FLIGHT_DATE	第一次飞行日期
	GENDER	性别
	FFP_TIER	会员卡级别
	WORK_CITY	工作地城市
	WORK_PROVINCE	工作地所在省份
	WORK_COUNTRY	工作地所在国家
	AGE	年龄
乘机信息	FLIGHT_COUNT	观测窗口内的飞行次数
	LOAD_TIME	观测窗口的结束时间
	LAST_TO_END	最后一次乘机时间至观测窗口结束时长
	AVG_DISCOUNT	平均折扣率
	SUM_YR	观测窗口的票价收入
	SEG_KM_SUM	观测窗口的总飞行公里数
	LAST_FLIGHT_DATE	末次飞行日期
	AVG_INTERVAL	平均乘机时间间隔
	MAX_INTERVAL	最大乘机间隔
积分信息	EXCHANGE_COUNT	积分兑换次数
	EP_SUM	总精英积分
	PROMOPTIVE_SUM	促销积分
	PARTNER_SUM	合作伙伴积分
	POINTS_SUM	总累计积分
	POINT_NOTFLIGHT	非乘机的积分变动次数
	BP_SUM	总基本积分

观测窗口：以过去某个时间点为结束时间，某一时间长度作为宽度，得到历史时间范围内的一个时间段。

表7-2　航空信息数据表

MEMBER_NO	FFP_DATE	FIRST_FLIGI	GENDE	FFP_TIER	WORK_CITY	WORK_PROVIN	WORK	AGE	LOAD_TIME	FLIGHT_COUNT	BP_SUM
289047040	2013/03/16	2013/04/28	男	6			US	56	2014/03/31	14	147 158
289053451	2012/06/26	2013/05/16	男	6	乌鲁木齐	新疆	CN	50	2014/03/31	65	112 582
289022508	2009/12/08	2010/02/05	男	5		北京	CN	34	2014/03/31	33	77 475
289004181	2009/12/10	2010/10/19	男	4	S.P.S	CORTES	HN	45	2014/03/31	6	76 027
289026513	2011/08/25	2011/08/25	男	6	乌鲁木齐	新疆	CN	47	2014/03/31	22	70 142
289027500	2012/09/26	2013/06/01	男	5	北京	北京	CN	36	2014/03/31	26	63 498
289058898	2010/12/27	2010/12/27	男	4	ARCADIA	CA	US	35	2014/03/31	5	62 810
289037374	2009/10/21	2009/10/21	男	4	广州	广东	CN	34	2014/03/31	4	60 484
289036013	2010/04/15	2013/06/02	女	6	广州	广东	CN	54	2014/03/31	25	59 357

（续）

MEMB-ER_NO	FFP_ DATE	FIRST_FLIGI	GENDE	FFP_ TIER	WORK_ CIT Y	WORK_ PROVIN	WORK	AGE	LOAD_ TIME	FLIGHT _COUNT	BP_ SUM
289046087	2007/01/26	2013/04/24	男	6	.	天津	CN	47	2014/03/31	36	55 562
289062045	2006/12/26	2013/04/17	女	5	长春市	吉林省	CN	55	2014/03/31	49	54 255
289061968	2011/08/15	2011/08/20	男	6	沈阳	辽宁	CN	41	2014/03/31	51	53 926
289022276	2009/08/27	2013/04/18	男	5	深圳	广东	CN	41	2014/03/31	62	49 224
289056049	2013/03/18	2013/07/28	男	4	Simi Valley		US	54	2014/03/31	12	49 121
289000500	2013/03/12	2013/04/01	男	5	北京	北京	CN	41	2014/03/31	65	46 618
289037025	2007/02/01	2011/08/22	男	6	昆明	云南	CN	57	2014/03/31	28	45 531
289029053	2004/12/18	2005/05/06	男	4			CN	46	2014/03/31	6	41 872
289048589	2008/08/15	2008/08/15	男	5	NUMAZU		CN	60	2014/03/31	15	41 610
289005632	2011/08/09	2011/08/09	男	5	南阳县	河南	CN	47	2014/03/31	6	40 726
289041886	2011/11/23	2013/09/17	女	5	温州	浙江	CN	42	2014/03/31	7	40 589
289049670	2010/04/18	2010/04/18	男	5	广州	广东	CN	39	2014/03/31	35	39 973
289020872	2008/06/22	2013/06/30	男	6	.	北京	CN	47	2014/03/31	33	39 737
289021001	2008/03/09	2013/07/10	男	6			CN	47	2014/03/31	40	39 584
289041371	2011/10/15	2013/09/04	男	6	武汉	湖北	CN	56	2014/03/31	30	38 089
289062046	2007/10/19	2007/10/19	男	5	上海	上海	CN	39	2014/03/31	48	37 188
289037246	2007/08/30	2013/04/18	男	6	贵阳	贵州	CN	47	2014/03/31	40	36 471
289045852	2006/08/16	2006/11/08	男	4	ARCADIA	CA	US	69	2014/03/31	8	35 707

数据详见：示例程序 /data/air_data.csv

7.2 分析方法与过程

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。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[15]。

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[15]。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 7-3），记为 LRFMC 模型。

表7-3　指标含义

模　型	L	R	F	M	C
航空公司 LRFMC 模型	会员入会时间距观测窗口结束的月数	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，如图 7-1 所示 [16]（它是依据属性的平均值进行划分，其中大于平均值的表示为↑，小于平均值的表示为↓），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。

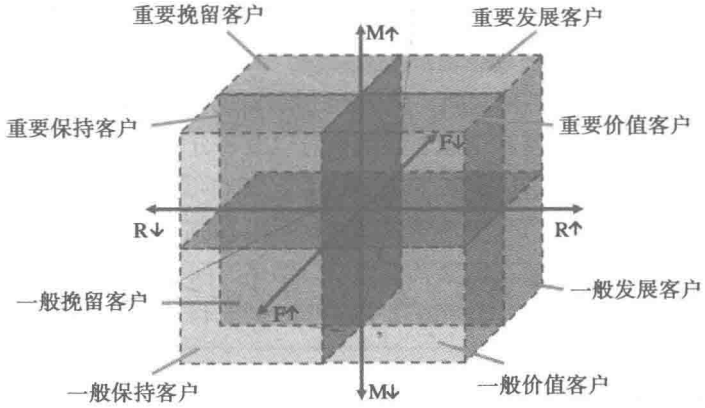

图 7-1　RFM 模型分析

本案例航空客户价值分析的总体流程如图 7-2 所示。

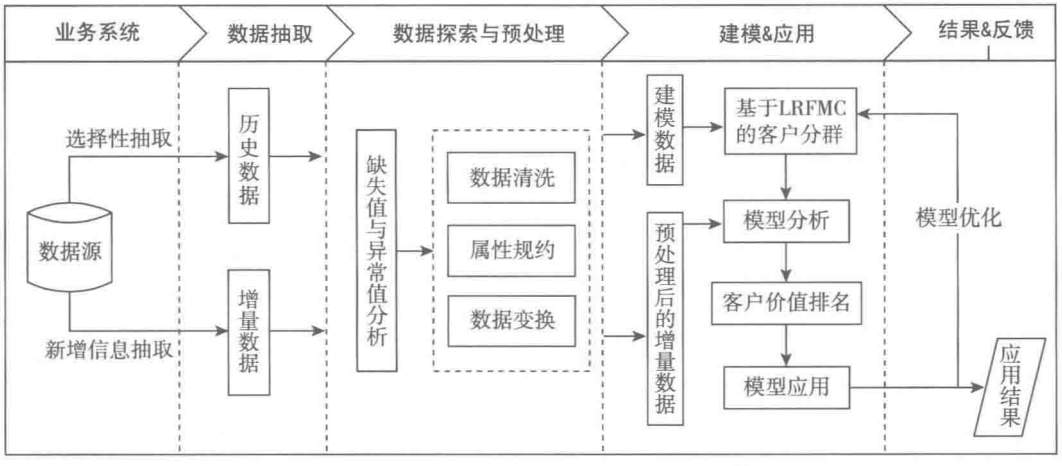

图 7-2　航空客运数据挖掘建模总体流程

航空客运信息挖掘主要包括以下步骤。

1）从航空公司的数据源中进行选择性抽取与新增数据抽取分别形成历史数据和增量数据。

2）对步骤 1）中形成的两个数据集进行数据探索分析与预处理，包括数据缺失值与异常值的探索分析，数据的属性规约、清洗和变换。

3）利用步骤 2）中形成的已完成数据预处理的建模数据，基于旅客价值 LRFMC 模型进行客户分群，对各个客户群进行特征分析，识别出有价值的客户。

4）针对模型结果得到不同价值的客户，采用不同的营销手段，提供定制化的服务。

7.2.1 数据抽取

以 2014-03-31 为结束时间，选取宽度为两年的时间段作为分析观测窗口，抽取观测窗口内有乘机记录的所有客户的详细数据形成历史数据。对于后续新增的客户详细信息，以后续新增数据中最新的时间点作为结束时间，采用上述同样的方法进行抽取，形成增量数据。

从航空公司系统内的客户基本信息、乘机信息以及积分信息等详细数据中，根据末次飞行日期（LAST_FLIGHT_DATE），抽取 2012-04-01 至 2014-03-31 内所有乘客的详细数据，总共有 62 988 条记录。其中包含了会员卡号、入会时间、性别、年龄、会员卡级别、工作地城市、工作地所在省份、工作地所在国家、观测窗口结束时间、观测窗口乘机积分、飞行公里数、飞行次数、飞行时间、乘机时间间隔和平均折扣率等 44 个属性。

7.2.2 数据探索分析

本案例的探索分析是对数据进行缺失值分析与异常值分析，分析出数据的规律以及异常值。通过对数据观察发现原始数据中存在票价为空值，票价最小值为 0、折扣率最小值为 0、总飞行公里数大于 0 的记录。票价为空值的数据可能是客户不存在乘机记录造成，其他的数据可能是客户乘坐 0 折机票或者积分兑换产生的。

查找每列属性观测值中空值个数、最大值、最小值的 Python 代码如代码清单 7-1 所示。

代码清单7-1 数据探索分析代码

```
#-*- coding: utf-8 -*-
#对数据进行基本的探索
#返回缺失值个数以及最大最小值

import pandas as pd

datafile= '../data/air_data.csv' #航空原始数据,第一行为属性标签
resultfile = '../tmp/explore.xls' #数据探索结果表

data = pd.read_csv(datafile, encoding = 'utf-8') #读取原始数据，指定UTF-8编码（需要用
    文本编辑器将数据转换为UTF-8编码）
```

```
explore = data.describe(percentiles = [], include = 'all').T #包括对数据的基本描述，
    percentiles参数是指定计算多少的分位数表（如1/4分位数、中位数等）；T是转置，转置后更方便查阅
explore['null'] = len(data)-explore['count'] #describe()函数自动计算非空值数，需要手
    动计算空值数

explore = explore[['null', 'max', 'min']]
explore.columns = [u'空值数', u'最大值', u'最小值'] #表头重命名
'''这里只选取部分探索结果。
describe()函数自动计算的字段有count（非空值数）、unique（唯一值数）、top（频数最高者）、freq
    （最高频数）、mean（平均值）、std（方差）、min（最小值）、50%（中位数）、max（最大值）'''

explore.to_excel(resultfile) #导出结果
```

代码详见：示例程序 /code/data_explore.py

根据上面的代码得到的探索结果见表 7-4。

表7-4　数据探索分析结果表

属性名称	空值记录数	最　大　值	最　小　值
SUM_YR_1	551	239 560	0
SUM_YR_2	138	234 188	0
…	…	…	…
SEG_KM_SUM	0	580 717	368
AVG_DISCOUNT	0	1.5	0

7.2.3　数据预处理

本案例主要采用数据清洗、属性规约与数据变换的预处理方法。

1. 数据清洗

通过数据探索分析，发现数据中存在缺失值，票价最小值为 0、折扣率最小值为 0、总飞行公里数大于 0 的记录。由于原始数据量大，这类数据所占比例较小，对于问题影响不大，因此对其进行丢弃处理。具体处理方法如下。

- 丢弃票价为空的记录。
- 丢弃票价为 0、平均折扣率不为 0、总飞行公里数大于 0 的记录。

使用 Pandas 对满足清洗条件的数据进行丢弃，处理方法：满足清洗条件的一行数据全部丢弃，其代码如代码清单 7-2 所示。

代码清单7-2　数据清洗代码

```
#-*- coding: utf-8 -*-
#数据清洗，过滤掉不符合规则的数据

import pandas as pd
```

```
datafile= '../data/air_data.csv' #航空原始数据,第一行为属性标签
cleanedfile = '../tmp/data_cleaned.csv' #数据清洗后保存的文件

data = pd.read_csv(datafile,encoding='utf-8') #读取原始数据，指定UTF-8编码（需要用文本
    编辑器将数据装换为UTF-8编码）

data = data[data['SUM_YR_1'].notnull()*data['SUM_YR_2'].notnull()] #票价非空值才保留

#只保留票价非零的，或者平均折扣率与总飞行公里数同时为0的记录。
index1 = data['SUM_YR_1'] != 0
index2 = data['SUM_YR_2'] != 0
index3 = (data['SEG_KM_SUM'] == 0) & (data['avg_discount'] == 0) #该规则是“与”
data = data[index1 | index2 | index3] #该规则是“或”

data.to_excel(cleanedfile) #导出结果
```

代码详见：示例程序 /code/data_clean.py

2. 属性规约

原始数据中属性太多，根据航空公司客户价值 LRFMC 模型，选择与 LRFMC 指标相关的 6 个属性：FFP_DATE、LOAD_TIME、FLIGHT_COUNT、AVG_DISCOUNT、SEG_KM_SUM、LAST_TO_END。删除与其不相关、弱相关或冗余的属性，例如，会员卡号、性别、工作地城市、工作地所在省份、工作地所在国家和年龄等属性。经过属性选择后的数据集，见表 7-5。

表7-5 属性选择后的数据集

LOAD_TIME	FFP_DATE	LAST_TO_END	FLIGHT_COUNT	SEG_KM_SUM	AVG_DISCOUNT
2014/3/31	2013/3/16	23	14	126 850	1.02
2014/3/31	2012/6/26	6	65	184 730	0.76
2014/3/31	2009/12/8	2	33	60 387	1.27
2014/3/31	2009/12/10	123	6	62 259	1.02
2014/3/31	2011/8/25	14	22	54 730	1.36
2014/3/31	2012/9/26	23	26	50 024	1.29
2014/3/31	2010/12/27	77	5	61 160	0.94
2014/3/31	2009/10/21	67	4	48 928	1.05
2014/3/31	2010/4/15	11	25	43 499	1.33
2014/3/31	2007/1/26	22	36	68 760	0.88
2014/3/31	2006/12/26	4	49	64 070	0.91
2014/3/31	2011/8/15	22	51	79 538	0.74
2014/3/31	2009/8/27	2	62	91 011	0.67
2014/3/31	2013/3/18	9	12	69 857	0.79

（续）

LOAD_TIME	FFP_DATE	LAST_TO_END	FLIGHT_COUNT	SEG_KM_SUM	AVG_DISCOUNT
2014/3/31	2013/3/12	2	65	75 026	0.69
2014/3/31	2007/2/1	13	28	50 884	0.86
2014/3/31	2004/12/18	56	6	73 392	0.66
2014/3/31	2008/8/15	23	15	36 132	1.07
2014/3/31	2011/8/9	48	6	55 242	0.79
2014/3/31	2011/11/23	36	7	44 175	0.89

3. 数据变换

数据变换是将数据转换成“适当的”格式，以适应挖掘任务及算法的需要。本案例中主要采用的数据变换方式为属性构造和数据标准化。

由于原始数据中并没有直接给出 LRFMC 五个指标，需要通过原始数据提取这五个指标，具体的计算方式如下。

（1）L = LOAD_TIME − FFP_DATE

会员入会时间距观测窗口结束的月数＝观测窗口的结束时间－入会时间 [单位：月]

（2）R = LAST_TO_END

客户最近一次乘坐公司飞机距观测窗口结束的月数＝最后一次乘机时间至观察窗口末端时长 [单位：月]

（3）F = FLIGHT_COUNT

客户在观测窗口内乘坐公司飞机的次数＝观测窗口的飞行次数 [单位：次]

（4）M = SEG_KM_SUM

客户在观测时间内在公司累计的飞行里程＝观测窗口的总飞行公里数 [单位：公里]

（5）C = AVG_DISCOUNT

客户在观测时间内乘坐舱位所对应的折扣系数的平均值＝平均折扣率 [单位：无]

5 个指标的数据提取后，对每个指标数据分布情况进行分析，其数据的取值范围见表 7-6。从表中数据可以发现，5 个指标的取值范围数据差异较大，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。

表7-6　LRFMC指标取值范围

属 性 名 称	L	R	F	M	C
最小值	12.23	0.03	2	368	0.14
最大值	114.63	24.37	213	580 717	1.5

标准差标准化处理的 Python 代码如代码清单 7-3 所示，datafile 为输入数据文件，zscoredata 为标准差标准化后数据集。

代码清单7-3 标准差标准化

```
#-*- coding: utf-8 -*-
#标准差标准化

import pandas as pd

datafile = '../data/zscoredata.xls' #需要进行标准化的数据文件;
zscoredfile = '../tmp/zscoreddata.xls' #标准差化后的数据存储路径文件;

#标准化处理
data = pd.read_excel(datafile)
data = (data - data.mean(axis = 0))/(data.std(axis = 0)) #简洁的语句实现了标准化变
    换，类似地可以实现任何想要的变换。
data.columns=['Z'+i for i in data.columns] #表头重命名。

data.to_excel(zscoredfile, index = False) #数据写入
```

代码详见：示例程序 /code/zscore_data.py

标准差标准化处理后，形成 ZL、ZR、ZF、ZM、ZC 5 个属性的数据，如表 7-7 所示。

表7-7 标准化处理后的数据集

ZL	ZR	ZF	ZM	ZC
1.690	0.140	−0.636	0.069	−0.337
1.690	−0.322	0.852	0.844	−0.554
1.682	−0.488	−0.211	0.159	−1.095
1.534	−0.785	0.002	0.273	−1.149
0.890	−0.427	−0.636	−0.685	1.232
−0.233	−0.691	-0.636	−0.604	−0.391
−0.497	1.996	−0.707	−0.662	−1.311
−0.869	−0.268	−0.281	−0.262	3.396
−1.075	0.025	−0.423	−0.521	0.150
1.907	−0.884	2.979	2.130	0.366
0.478	−0.565	0.852	−0.068	−0.662
0.469	−0.939	0.073	0.104	−0.013
0.469	−0.185	-0.140	−0.220	−0.932
0.453	1.517	0.073	−0.301	3.288
0.369	0.747	−0.636	−0.626	−0.283
0.312	−0.896	0.498	0.954	−0.500
−0.026	−0.681	0.073	0.325	0.366
−0.051	2.723	−0.636	−0.749	0.799
−0.092	2.879	−0.707	−0.734	−0.662
−0.150	−0.521	1.278	1.392	1.124

数据详见：示例程序 /data/zscoreddata.xls

7.2.4　模型构建

客户价值分析模型构建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。

1. 客户聚类

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）。

K-Means 聚类算法位于 Scikit-Learn 库下的聚类子库（sklearn.cluster），代码如代码清单 7-4 所示，输入数据集为 inputfile，聚类类别数为 k = 5。

代码清单7-4　K-Means聚类算法

```
#-*- coding: utf-8 -*-
#K-Means聚类算法

import pandas as pd
from sklearn.cluster import KMeans #导入K均值聚类算法

inputfile = '../tmp/zscoreddata.xls' #待聚类的数据文件
k = 5                          #需要进行的聚类类别数

#读取数据并进行聚类分析
data = pd.read_excel(inputfile) #读取数据

#调用k-means算法，进行聚类分析
kmodel = KMeans(n_clusters = k, n_jobs = 4) #n_jobs是并行数，一般等于CPU数较好
kmodel.fit(data) #训练模型

kmodel.cluster_centers_ #查看聚类中心
kmodel.labels_ #查看各样本对应的类别
```

代码详见：示例程序 /code/KMeans_cluster.py

对数据进行聚类分群的结果如表 7-8 所示。

表7-8　客户聚类结果

聚类类别	聚类个数	聚类中心				
		ZL	ZR	ZF	ZM	ZC
客户群 1	5 337	0.483	−0.799	2.483	2.424	0.308
客户群 2	15 735	1.160	−0.377	−0.087	−0.095	−0.158
客户群 3	12 130	−0.314	1.686	−0.574	−0.537	−0.171
客户群 4	24 644	−0.701	−0.415	−0.161	−0.165	−0.255
客户群 5	4 198	0.057	−0.006	−0.227	−0.230	2.191

注：由于 K-Means 聚类是随机选择类标号，因此重复此实验得到结果中的类标号可能与此不同；另外，由于算法的精度问题，重复实验得到的聚类中心也可能略有不同。

2. 客户价值分析

针对聚类结果进行特征分析，如图 7-3 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 7-9 所示。

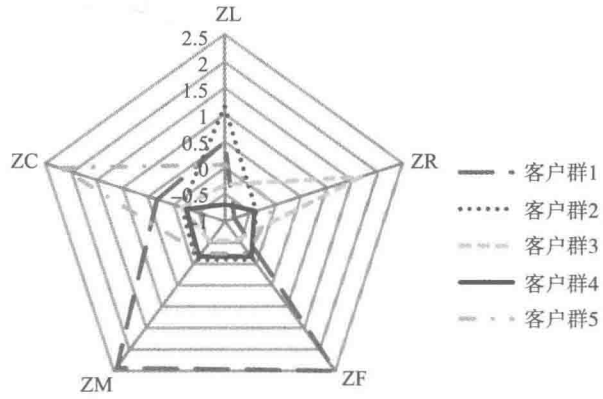

图 7-3 客户群特征分析图

表7-9 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

注：正常字体表示最大值、加粗字体表示次大值、斜体字体表示最小值、带下划线的字体表示次小值。

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。他们之间的区别如图 7-4 所示，其中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。

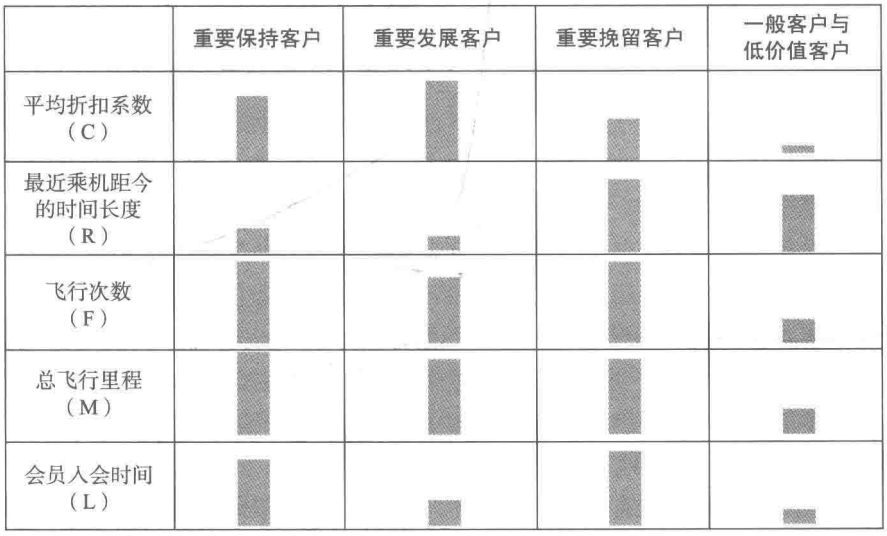

图 7-4　客户类别的特征分析

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 7-10 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7-10　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

（续）

客 户 群	排 名	排名含义
客户群 2	3	重要挽留用户
客户群 4	4	一般客户
客户群 3	5	低价值客户

本模型采用历史数据进行建模，随着时间的变化，分析数据的观测窗口也在变换。因此，对于新增客户详细信息，考虑业务的实际情况，该模型建议每一个月运行一次，对其新增客户信息通过聚类中心进行判断，同时对本次新增客户的特征进行分析。如果增量数据的实际情况与判断结果差异大，需要业务部门重点关注，查看变化大的原因以及确认模型的稳定性。如果模型稳定性变化大，需要重新训练模型进行调整。目前模型进行重新训练的时间没有统一标准，大部分情况都是根据经验来决定。根据经验建议：每隔半年训练一次模型比较合适。

3. 模型应用

根据对各个客户群进行特征分析，采取下面的一些营销手段和策略，为航空公司的价值客户群管理提供参考。

（1）会员的升级与保级

航空公司的会员可以分为白金卡会员、金卡会员、银卡会员、普通卡会员，其中非普通卡会员可以统称为航空公司的精英会员。虽然各个航空公司都有自己的特点和规定，但会员制的管理方法是大同小异的。成为精英会员一般都是要求在一定时间内（如一年）积累一定的飞行里程或航段，达到这种要求后就会在有效期内（通常为两年）成为精英会员，并享受相应的高级别服务。有效期快结束时，根据相关评价方法确定客户是否有资格继续作为精英会员，然后对该客户进行相应地升级或降级。

然而，由于许多客户并没有意识到或根本不了解会员升级或保级的时间与要求（相关的文件说明往往复杂且不易理解），经常在评价期过后才发现自己其实只差一点就可以实现升级或保级，却错过了机会，使之前的里程积累白白损失。同时，这种认知还可能导致客户的不满，干脆放弃在本公司的消费。

因此，航空公司可以在对会员升级或保级进行评价的时间点之前，对那些接近但尚未达到要求的较高消费客户进行适当提醒甚至采取一些促销活动，刺激他们通过消费达到相应标准。这样既可以获得收益，同时也提高了客户的满意度，增加了公司的精英会员。

（2）首次兑换

航空公司常旅客计划中最能够吸引客户的内容就是客户可以通过消费积累的里程来兑换免票或免费升舱等。各个航空公司都有一个首次兑换标准，也就是当客户的里程或航段积累到一定程度时才可以实现第一次兑换，这个标准会高于正常的里程兑换标准。但是很多公司的里程积累随着时间会进行一定地削减，例如有的公司会在年末对该年积累的里程进行折半

处理。这样会导致许多不了解情况的会员白白损失自己好不容易积累的里程，甚至总是难以实现首次兑换。同样，这也会引起客户的不满或流失。可以采取的措施是从数据库中提取出接近但尚未达到首次兑换标准的会员，对他们进行提醒或促销，使他们通过消费达到标准。一旦实现了首次兑换，客户在本公司进行再次消费兑换就比在其他公司进行兑换要容易许多，在一定程度上等于提高了转移的成本。另外，在一些特殊的时间点（如里程折半的时间点）之前可以给客户一些提醒，这样可以增加客户的满意度。

（3）交叉销售

通过发行联名卡等与非航空类企业的合作，使客户在其他企业的消费过程中获得本公司的积分，增强与公司的联系，提高他们的忠诚度。例如，可以查看重要客户在非航空类合作伙伴处的里程积累情况，找出他们习惯的里程积累方式（是否经常在合作伙伴处消费、更喜欢消费哪些类型合作伙伴的产品），对他们进行相应促销。

客户识别期和发展期为客户关系打下基石，但是这两个时期带来的客户关系是短暂的、不稳定的。企业要获取长期的利润，必须具有稳定的、高质量的客户。保持客户对于企业是至关重要的，不仅因为争取一个新客户的成本远远高于维持老客户的成本，更重要的是客户流失会造成公司收益的直接损失。因此，在这一时期，航空公司应该努力维系客户关系，使之处于较高的水准，最大化生命周期内公司与客户的互动价值，并使这样的高水平尽可能延长。对于这一阶段的客户，主要应该通过提供优质的服务产品和提高服务水平来提高客户的满意度。通过对旅客数据库的数据挖掘、进行客户细分，可以获得重要保持客户的名单。这类客户一般所乘航班的平均折扣率（C）较高，最近乘坐过本公司航班（R 低）、乘坐的频率（F）或里程（M）也较高。他们是航空公司的价值客户，是最理想的客户类型，对航空公司的贡献最大，所占比例却比较小。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。

7.3　上机实验

1. 实验目的

- 了解 K-Means 聚类算法在客户价值分析实例中的应用。
- 利用 Pandas 快速实现数据 z-score（标准差）标准化以及用 Scikit-Learn 的聚类库实现 K-Means 聚类。

2. 实验内容

依据航空公司客户价值分析的 LRFMC 模型提取客户信息的 LRFMC 指标。对其进行标准差标准化并保存后，采用 K-Means 算法完成客户的聚类，分析每类的客户特征，从而获得每类的客户价值。

- 利用 Pandas 程序，读入 LRFMC 指标文件，分别计算各个指标的均值和标准差，使用标准差标准化公式完成 LRFMC 指标的标准化，并将标准化后的数据进行保存。
- 编写 Python 程序，完成客户的 K-Means 聚类，获得聚类中心与类标号。输出聚类中心的特征图，并统计每个类别的客户数。

3. 实验方法与步骤

实验一

对 L、R、F、M、C 五个指标进行 z-score（标准差）标准化。

1）启动 Python 并导入 Pandas，使用 read_excel() 函数将待标准差标准化的数据“上机实验 /data/zscoredata.xls”读入到 Python 中。

2）使用 mean() 与 std() 函数，获得 L、R、F、M、C 五个指标的平均值与标准差。

3）根据 z-score（标准差）标准化公式 $z_{ij} = (x_{ij} - x_i)/s_i$，其中 z_{ij} 是标准化后的变量值；x_{ij} 是实际变量值，x_i 为变量的算术平均值，s_i 是变量的标准差，进行标准差标准化。

实验二

1）使用 read_excel 函数将航空数据预处理后的数据读入 Python 工作空间，截取最后 5 列数据作为 K-Means 算法的输入数据。

2）调用 KMeans 函数对 1）中的数据进行聚类，得到聚类标号和聚类中心点。

3）根据聚类标号统计每个类别的客户数，同时根据聚类中心点向量画出客户聚类中心向量图并保存。

4. 思考与实验总结

1）Scikit-Learn 中 KMeans 函数中的初始聚类中心可以使用什么算法得到？默认是什么算法？

2）使用不同的预处理对原始数据进行变换，再使用 K-Means 算法进行聚类，对比聚类结果，分析不同数据预处理对 K-Means 算法的影响。

7.4 拓展思考

本章主要针对客户价值进行分析，对客户流失并没有提出具体的分析。由于在航空客户关系管理中客户流失的问题未被重视，故对航空公司造成了巨大的损害。客户流失对利润增长造成的负面影响非常大，仅次于公司规模、市场占有率和单位成本等因素的影响。客户与航空公司之间的关系越长久，给航空公司带来的利润就会越高。所以流失一个客户，比获得一个新客户对公司的损失更大。因为要获得新客户，需要在销售、市场、广告和人员工资上花费很多，并且大多数新客户产生的利润不如那些流失的老客户多。

因此，在国内航空市场竞争日益激烈的背景下，航空公司在客户流失方面应该引起足够的重视。如何改善流失问题，继而提高客户满意度、忠诚度是航空公司维护自身市场并面对

激烈竞争的一件大事，客户流失分析将成为帮助航空公司开展持续改进活动的指南。

客户流失分析可以针对目前老客户进行分类预测。针对航空公司客户信息数据（见表 7-2），可以进行老客户以及客户类型的定义（其中将飞行次数大于 6 次的客户定义为老客户，已流失客户定义为：第二年飞行次数与第一年飞行次数比例小于 50% 的客户；准流失客户定义为：第二年飞行次数与第一年飞行次数比例在［50%，90%）内的客户；未流失客户定义为：第二年飞行次数与第一年飞行次数比例大于 90% 的客户）。同时，需要选取客户信息中的关键属性，如会员卡级别、客户类型（流失、准流失、未流失）、平均乘机时间间隔、平均折扣率、积分兑换次数、非乘机积分总和、单位里程票价和单位里程积分等。随机选取数据的 80% 作为分类的训练样本，剩余的 20% 作为测试样本。构建客户的流失模型，运用模型预测未来客户的类别归属（未流失、准流失或已流失）。

7.5　小结

本章结合航空公司客户价值分析的案例，重点介绍了数据挖掘算法中 K-Means 聚类算法在实际案例中的应用。针对客户价值识别传统的 RFM 模型的不足，采用 K-Means 算法进行分析，并详细地描述了数据挖掘的整个过程，对其相应的算法给出了 Python 上机实验步骤。