



航空公司客户价值分析

张敏

1	背景与目标
2	数据预处理
3	特征构造
4	K-Means客户分群
5	小结

1. 行业内竞争

民航的竞争除了三大航空公司之间的竞争之外，还将加入新崛起的各类小型航空公司、民营航空公司，甚至国外航空巨头。航空产品生产过剩，产品同质化特征愈加明显，于是航空公司从价格、服务间的竞争逐渐转向对客户的竞争。



2. 行业外竞争

随着高铁、动车等铁路运输的兴建，航空公司受到巨大冲击。



客户营销战略倡导者Jay & Adam Curry从国外数百家公司进行了客户营销实施的经验中提炼了如下经验。

- 公司收入的80%来自顶端的20%的客户。
- 20%的客户其利润率100%。
- 90%以上的收入来自现有客户。
- 大部分的营销预算经常被用在非现有客户上。
- 5%至30%的客户在客户金字塔中具有升级潜力。
- 客户金字塔中客户升级2%，意味着销售收入增加10%，利润增加50%。

这些经验也许并不完全准确，但是它揭示了新时代客户分化的趋势，也说明了对客户价值分析的迫切性和必要性。

航空公司数据特征说明

- 目前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。
- 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，44个特征，总共62988条记录。数据特征及其说明如右表所示。

	特征名称	特征说明
客户基本信息	MEMBER_NO	会员卡号
	FFP_DATE	入会时间
	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

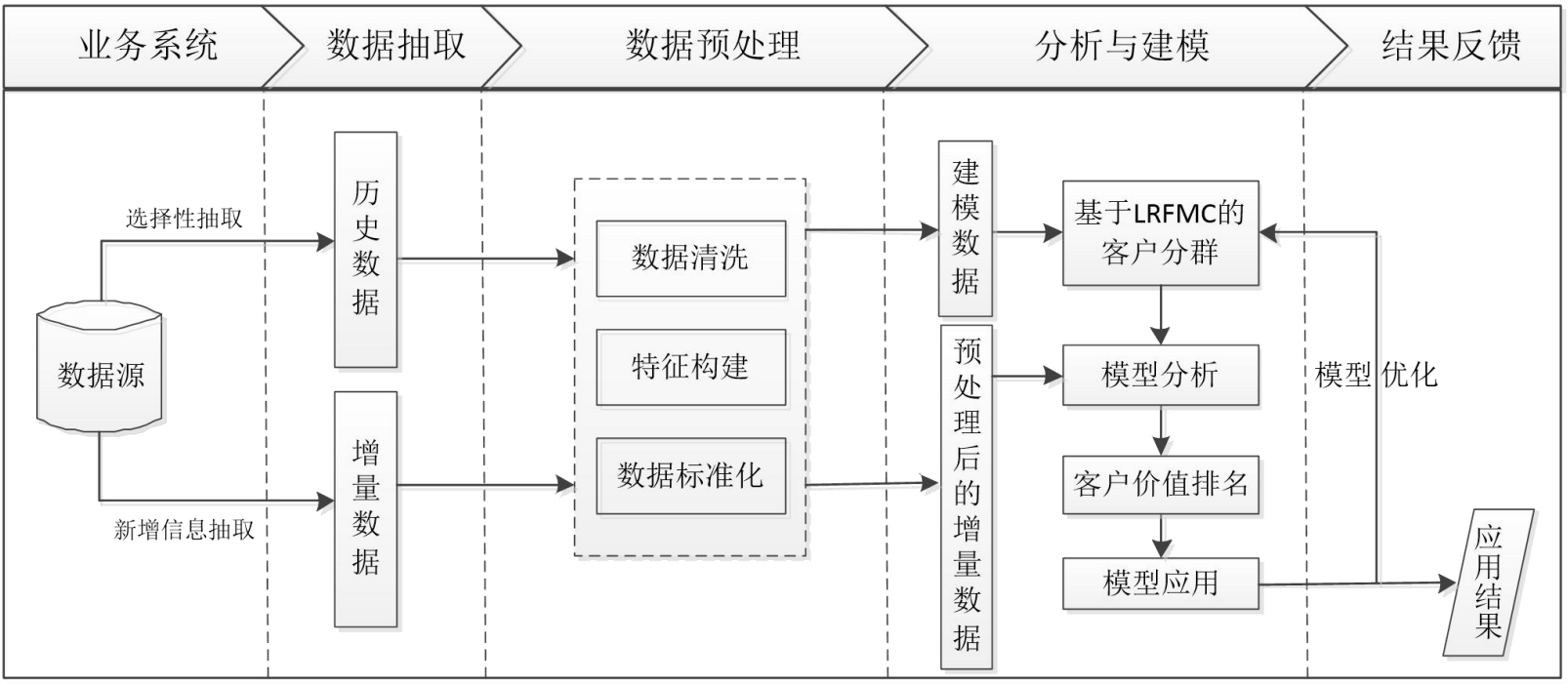
表 名	特征名称	特征说明
乘机信息	FLIGHT_COUNT	观测窗口内的飞行次数
	LOAD_TIME	观测窗口的结束时间
	LAST_TO_END	最后一次乘机时间至观测窗口结束时长
	AVG_DISCOUNT	平均折扣率
	SUM_YR	观测窗口的票价收入
	SEG_KM_SUM	观测窗口的总飞行公里数
	LAST_FLIGHT_DATE	末次飞行日期
	AVG_INTERVAL	平均乘机时间间隔
	MAX_INTERVAL	最大乘机间隔
积分信息	EXCHANGE_COUNT	积分兑换次数
	EP_SUM	总精英积分
	PROMOPTIVE_SUM	促销积分
	PARTNER_SUM	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

结合目前航空公司的数据情况，可以实现以下目标。

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，比较不同类别客户的客户价值。
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。

1	背景与目标
2	数据预处理
3	特征构造
4	K-Means客户分群
5	小结

航空客户价值分析项目的总体流程如图所示。



航空公司客户原始数据存在少量的缺失值和异常值，需要清洗后才能用于分析。

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。

处理方法：丢弃票价为空的记录。

➤ 其他的数据可能是客户乘坐0折机票或者积分兑换造成。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。

处理方法：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。

1	背景与目标
2	数据预处理
3	特征构造
4	K-Means客户分群
5	小结

原始数据中包含40多个特征，利用这些特征做些什么呢？我们又该从哪些角度出发呢？

1. RFM模型介绍

本项目的目标是客户价值分析，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RFM模型。

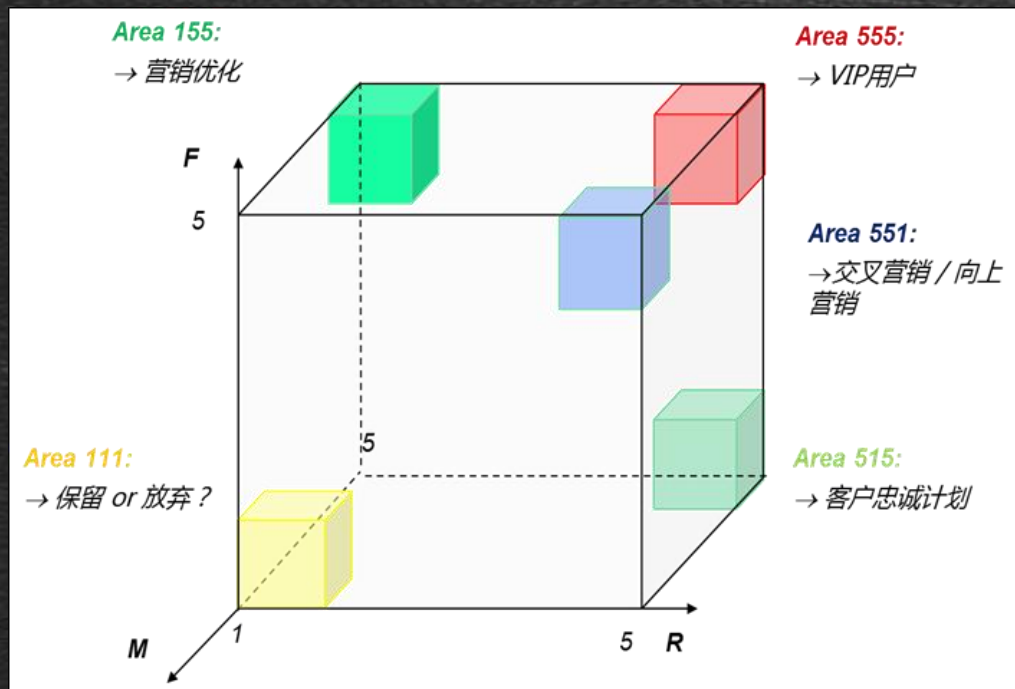
- R (Recency) 指的是最近一次消费时间与截止时间的间隔。
- F (Frequency) 指顾客在某段时间内所消费的次数。
- M (Monetary) 指顾客在某段时间内所消费的金额。

2. RFM模型结果解读

RFM模型包括三个特征，使用三维坐标系进行展示，如图所示。

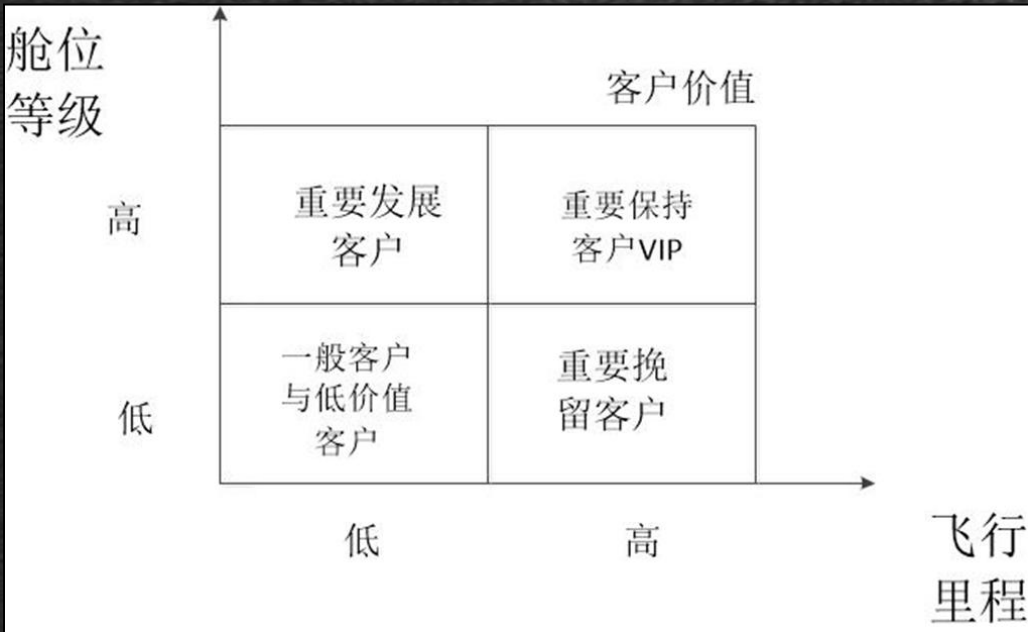
- X轴表示Recency,
- Y轴表示Frequency,
- Z轴表示Monetary,

每个轴一般会分成5级表示程度，1为最小，5为最大。



3. 传统RFM模型在航空行业的缺陷

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，因此这个特征并不适合用于航空公司的客户价值分析。



4. 航空客户价值分析的LRFMC模型

本项目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两个特征代替消费金额。此外，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特征。

本项目将客户关系长度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作为航空公司识别客户价值的关键特征（如表 3 2所示），记为LRFMC模型。

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

完成五个特征的构建以后，对每个特征数据分布情况进行分析，其数据的取值范围如表所示。从表中数据可以发现，五个特征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，需要对数据做标准化处理。

特征名称	L	R	F	M	C
最小值	12.17	0.03	2	368	0.14
最大值	114.57	24.37	213	580717	1.5

L、R、F、M和C五个特征的数据示例，上图为原始数据，下图为标准差标准化处理后的数据。

LOAD_TIME	FFP_DATE	LAST_TO_END	FLIGHT_COUNT	SEG_KM_SUM	AVG_DISTANCE
2014/3/31	2013/3/16	23	14	126850	1.02
2014/3/31	2012/6/26	6	65	184730	0.76
2014/3/31	2009/12/8	2	33	60387	1.27
2014/3/31	2009/12/10	123	6	62259	1.02
2014/3/31	2011/8/25	14	22	54730	1.36

L	R	F	M	C
1.44	-0.95	14.03	26.76	1.30
1.31	-0.91	9.07	13.13	2.87
1.33	-0.89	8.72	12.65	2.88
0.66	-0.42	0.78	12.54	1.99
0.39	-0.92	9.92	13.90	1.34

1	背景与目标
2	数据预处理
3	特征构造
4	K-Means客户分群
5	小结

1. 基本概念

K-Means聚类算法是一种基于质心的划分方法，输入聚类个数 k ，以及包含 n 个数据对象的数据库，输出满足误差平方和最小标准的 k 个聚类。算法步骤如下。

- 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。
- 分别计算每个样本到各个聚类质心的距离，将样本分配到距离最近的那个聚类中心类别中。
- 所有样本分配完成后，重新计算 k 个聚类的中心。
- 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(2)，否则转(5)。
- 当质心不发生变化时停止并输出聚类结果。

2. 数据类型

K-Means聚类算法是在数值类型数据的基础上进行研究，然而数据分析的样本复杂多样，因此要求不仅能够对特征为数值类型的数据进行分析，还要适应数据类型的变化，对不同特征做不同变换，以满足算法的要求。



3. kmeans函数及其参数介绍

- sklearn的cluster模块提供了KMeans函数构建K-Means聚类模型，其基本语法如下。

```
sklearn.cluster.KMeans(n_clusters=8, init='k-means++', n_init=10, max_iter=300,  
tol=0.0001, precompute_distances='auto', verbose=0, random_state=None,  
copy_x=True, n_jobs=1, algorithm='auto')
```

常用参数及其说明如表所示。

3. kmeans函数及其参数介绍

常用参数及其说明如表所示。

参数名称	说明
n_clusters	接收int。表示分类簇的数量。无默认。
max_iter	接收int。表示最大的迭代次数。默认为300。
n_init	接收int。表示算法的运行次数。默认为10。
init	接收特定string。kmeans++表示该初始化策略选择的初始均值向量相互之间都距离较远，它的效果较好；random表示从数据中随机选择K个样本作为初始均值向量；或者提供一个数组，数组的形状为（n_clusters,n_features），该数组作为初始均值向量。默认为kmeans++。
precompute_distances	接收boolean或者auto。表示是否提前计算好样本之间的距离，auto表示如果n_samples*n > 12million,则不提前计算。默认为auto。
tol	接收float。表示算法收敛的阈值。默认为0.0001。
n_jobs	接收int。表示任务使用的CPU数量。默认为1。
random_state	接收int。表示随机数生成器的种子。默认为None。
verbose	接收int。0表示不输出日志信息；1表示每隔一段时间打印一次日志信息，如果大于1，则打印日志信息更频繁。默认为0。

3. kmeans函数及其参数介绍

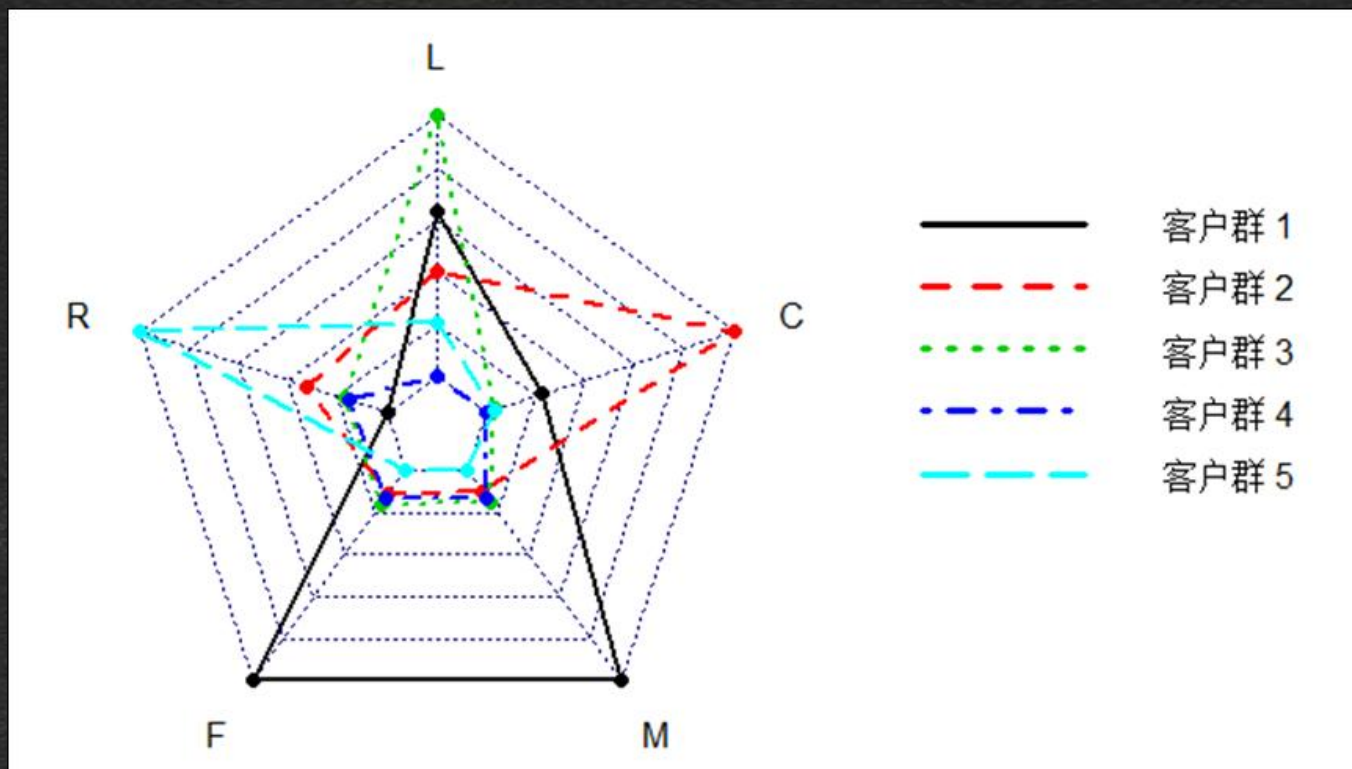
K-Means模型构建完成后可以通过属性查看不同的信息，如表所示。

属性	说明
cluster_centers_	返回ndarray。表示分类簇的均值向量。
labels_	返回ndarray。表示每个样本所属的簇的标记。
Inertia_	返回ndarray。表示每个样本距离它们各自最近簇中心之和。

对数据进行聚类分群的结果如表所示。

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





















针对聚类结果进行特征分析，如图所示。



结合业务分析，通过比较各个特征在群间的大小对某一个群的特征进行评价分析，从而总结出每个群的优势和弱势特征，具体结果如表所示。

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

基于特征描述，本项目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图所示。

	重要保持客户	重要发展客户	重要挽留客户	一般客户与低价值客户
平均折扣系数 (C)				
最近乘机距今的时间长度 (R)				
飞行次数 (F)				
总飞行里程 (M)				
会员入会时间 (L)				

1	背景与目标
2	数据预处理
3	特征构造
4	K-Means客户分群
5	小结

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。

- **会员的升级与保级：**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。
- **首次兑换：**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。
- **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。

本项目结合航空公司客户价值分析的案例，重点介绍了数据分析算法中K-Means聚类算法在客户价值分析中的应用。针对RFM客户价值分析模型的不足，使用K-Means算法构建了航空客户价值分析LRFMC模型，详细描述了数据分析的整个过程。





Thank you!