航空公司客户价值分析

谢佳标

背景

- 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指定优化的个性化服务方案，采取不同的营销策略，将优先营销资源集中于高价值客户，实现企业利润最大化目标。

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
	FIRST_FLIGHT_DATE	第一次飞行时间
	GENDER	性别
	FFP_TIER	会员卡级别
	WORK_CITY	工作地城市
	WORK_PROVINCE	工作地所在省市
	WORK_COUNTRY	工作地所在国家
	AGE	年龄

	属性名称	属性说明
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	最大乘机间隔

	属性名称	属性说明
积分信息	EXCHANGE_COUNT	积分兑换次数
	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

数据示例

MEMBER_NO	FFP_DATE	FIRST_FLIGHT_DATE	GENDER	FFP_TIER	WORK_CITY	WORK_PROVINCE	WORK_COUNTRY	AGE	LOAD_TIME	FLIGHT_COUNT	BP_SUM	EP_SUM_YR_1	EP_SUM_YR_2	SUM_YR_1	SUM_YR_2	SEG_KM_SUM
54993	2006/11/2	2008/12/24	男		6.	北京	CN	31	2014/3/31	210	505308	0	74460	239560	234188	580717
28065	2007/2/19	2007/8/3	男		6	北京	CN	42	2014/3/31	140	362480	0	41288	171483	167434	293678
55106	2007/2/1	2007/8/30	男		6.	北京	CN	40	2014/3/31	135	351159	0	39711	163618	164982	283712
21189	2008/8/22	2008/8/23	男		5Los Angeles	CA	US	64	2014/3/31	23	337314	0	34890	116350	125500	281336
39546	2009/4/10	2009/4/15	男		6贵阳	贵州	CN	48	2014/3/31	152	273844	0	42265	124560	130702	309928
56972	2008/2/10	2009/9/29	男		6广州	广东	CN	64	2014/3/31	92	313338	0	27323	112364	76946	294585
44924	2006/3/22	2006/3/29	男		6乌鲁木齐市	新疆	CN	46	2014/3/31	101	248864	0	37689	120500	114469	287042
22631	2010/4/9	2010/4/9	女		6温州市	浙江	CN	50	2014/3/31	73	301864	0	39834	82440	114971	287230
32197	2011/6/7	2011/7/1	男		5DRANCY		FR	50	2014/3/31	56	262958	0	31700	72596	87401	321489
31645	2010/7/5	2010/7/5	女		6温州	浙江	CN	43	2014/3/31	64	204855	0	47052	85258	60267	375074
58877	2010/11/18	2010/11/20	女		6PARIS	PARIS	FR	34	2014/3/31	43	298321	0	39018	69056	91581	262013
37994	2004/11/13	2004/12/2	男		6北京	.	CN	47	2014/3/31	145	256093	0	44539	92975	126821	271438
28012	2006/11/23	2007/11/18	男		5SAN MARINO	CA	US	58	2014/3/31	29	210269	0	35539	44750	53977	321529
54943	2006/10/25	2007/10/27	男		6深圳	广东	CN	47	2014/3/31	118	241614	0	26426	105466	119832	179514
57881	2010/2/1	2010/2/1	女		6广州	广东	CN	45	2014/3/31	50	289917	0	38028	68941	79076	270067
1254	2008/3/28	2008/4/5	男		4BOWLAND HEIGHTS	CALIFORNIA	US	63	2014/3/31	22	286164	0	23338	69300	54764	234721
8253	2010/7/15	2010/8/20	男		6乌鲁木齐	新疆	CN	48	2014/3/31	101	219995	0	23381	93840	93114	172231
58899	2010/11/10	2011/2/23	女		6PARIS		FR	50	2014/3/31	40	249882	0	31823	66239	63260	284160
26955	2006/4/6	2007/2/22	男		6乌鲁木齐市	新疆	CN	54	2014/3/31	64	215013	0	22036	99735	93006	169358
41616	2011/8/29	2011/10/22	男		6东莞	广东	CN	41	2014/3/31	38	191038	0	24656	60930	52316	332896
21501	2008/7/30	2008/11/21	男		6.	北京	CN	49	2014/3/31	106	220641	0	30493	69566	122763	167113
41281	2011/6/7	2011/6/9	男		6VECHEL	NORD BRABANT	AN		2014/3/31	23	255573	0	29947	46800	198224	214590
47229	2005/4/10	2005/4/10	男		6广州	广东	CN	69	2014/3/31	94	193169	0	35222	59169	74497	305250
28474	2010/4/13	2010/4/13	男		6	CA	US	41	2014/3/31	20	256337	0	24423	64258	59600	222380
58472	2010/2/14	2010/3/1	女		5		FR	48	2014/3/31	44	204801	0	30638	38510	75816	281837
13942	2010/10/14	2010/11/1	男		6PARIS	FRANCE	FR	39	2014/3/31	62	241719	0	32263	72806	83496	243674
45075	2007/2/1	2007/3/23	男		6湛江	广东	CN	46	2014/3/31	213	217809	0	20801	136769	96568	187917
47114	2005/1/15	2005/3/17	男		6.	北京	CN	47	2014/3/31	74	209810	0	20044	101398	83139	148685
54619	2006/1/7	2006/1/8	男		6广州	广东	CN	62	2014/3/31	101	209362	0	24873	94055	107896	159129
12349	2008/6/16	2008/6/27	男		6深圳	广东	CN	46	2014/3/31	87	199866	0	22439	83780	102594	149254
35883	2006/4/11	2007/4/18	男		6乌鲁木齐市	新疆	CN	55	2014/3/31	53	195075	0	16804	96945	69814	144076
56091	2004/11/25	2005/2/10	女		6广州市	广东省	CN	46	2014/3/31	95	197222	0	13133	134702	60366	149053
2137	2005/4/11	2005/5/3	男		6广州	广东	CN	69	2014/3/31	131	199716	0	29222	86875	89532	220948
27708	2006/3/20	2006/3/25	男		6深圳	广东	CN	45	2014/3/31	95	194326	0	17222	96692	75484	148227
28014	2006/12/1	2011/1/7	女		6Paris		FR	48	2014/3/31	47	218842	0	30870	63100	69298	285144
3539	2009/10/13	2009/10/13	男		4UPLAND	CALIFORNIA	US	53	2014/3/31	25	228851	0	29866	115502	48696	220389

挖掘目标

- 借助航空公司客户数据，对客户进行分类
- 对不同的用户类别进行特征分析，比较不同类客户的客户价值
- 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。

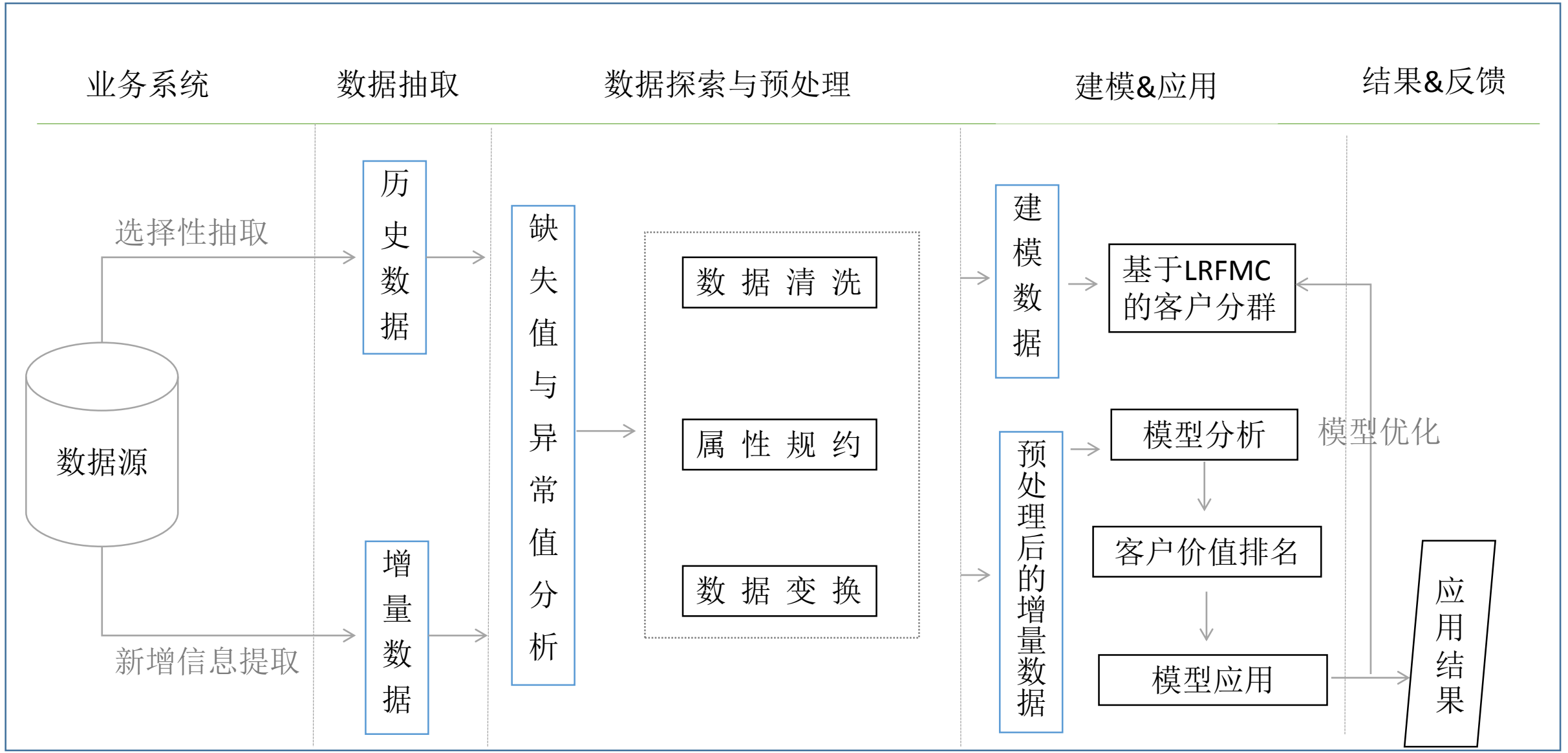
指标含义

- 本案例将客户入会时长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，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	客户在观测矿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

- 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五个指标进行K-Means聚类，识别最有价值客户。

航空客运数据挖掘建模总体流程



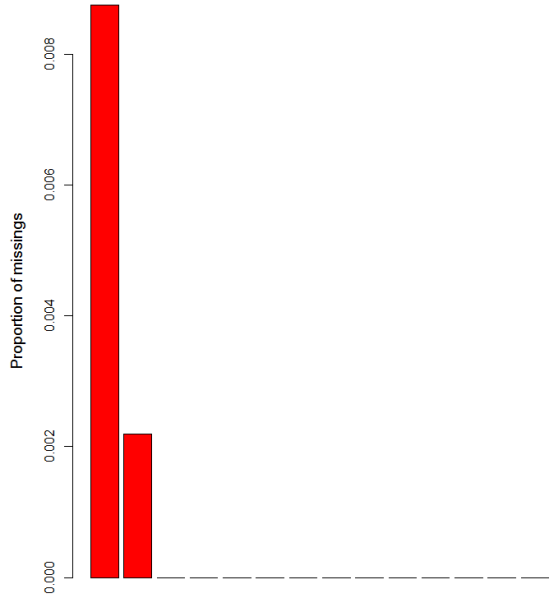
数据探索分析

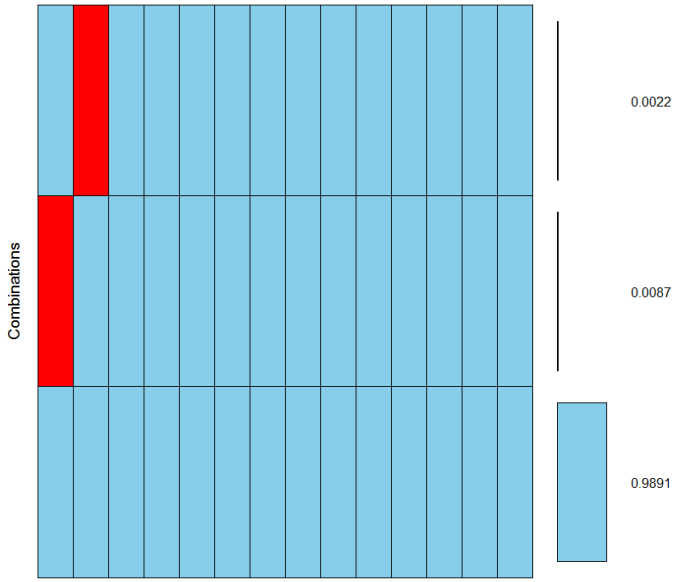
```
# 数据读取
datafile <- read.csv("air_data.csv")
# 选择需要探索分析的变量
col <- c(15:18,20:29) #去掉日期型变量
# 输出变量最值、缺失情况
summary(datafile[,col])
library(mice)
library(VIM)
md.pattern(datafile[,col])
aggr(datafile[,col],prob=F,number=T)
```

```
> md.pattern(datafile[,col])
      SEG_KM_SUM WEIGHTED_SEG_KM AVG_FLIGHT_COUNT AVG_BP_SUM BEGIN_TO_FIRST LAST_TO_END AVG_INTERVAL MAX_INTERVAL
62299          1                1                1          1              1           1            1            1
  551          1                1                1          1              1           1            1            1
  138          1                1                1          1              1           1            1            1
               0                0                0          0              0           0            0            0
      ADD_POINTS_SUM_YR_1 ADD_POINTS_SUM_YR_2 EXCHANGE_COUNT avg_discount SUM_YR_2 SUM_YR_1
62299                   1                   1              1            1       1       1   0
  551                   1                   1              1            1       1       0   1
  138                   1                   1              1            1       0       1   1
                        0                   0              0            0     138     551 689
```

```
> summary(datafile[,col])
    SUM_YR_1          SUM_YR_2         SEG_KM_SUM     WEIGHTED_SEG_KM  AVG_FLIGHT_COUNT
 Min.   :     0   Min.   :     0   Min.   :   368   Min.   :     0   Min.   : 0.2500
 1st Qu.:  1003   1st Qu.:   780   1st Qu.:  4747   1st Qu.:  3219   1st Qu.: 0.4286
 Median :  2800   Median :  2773   Median :  9994   Median :  6978   Median : 0.8750
 Mean   :  5355   Mean   :  5604   Mean   : 17124   Mean   : 12777   Mean   : 1.5422
 3rd Qu.:  6574   3rd Qu.:  6846   3rd Qu.: 21271   3rd Qu.: 15300   3rd Qu.: 1.8750
 Max.   :239560   Max.   :234188   Max.   :580717   Max.   :558440   Max.   :26.6250
 NA's   :551      NA's   :138
   AVG_BP_SUM      BEGIN_TO_FIRST   LAST_TO_END     AVG_INTERVAL     MAX_INTERVAL
 Min.   :    0.0   Min.   :  0.0   Min.   :  1.0   Min.   :  0.00   Min.   :  0
 1st Qu.:  336.0   1st Qu.:  9.0   1st Qu.: 29.0   1st Qu.: 23.37   1st Qu.: 79
 Median :  752.4   Median : 50.0   Median :108.0   Median : 44.67   Median :143
 Mean   : 1421.4   Mean   :120.1   Mean   :176.1   Mean   : 67.75   Mean   :166
 3rd Qu.: 1690.3   3rd Qu.:166.0   3rd Qu.:268.0   3rd Qu.: 82.00   3rd Qu.:228
 Max.   :63163.5   Max.   :729.0   Max.   :731.0   Max.   :728.00   Max.   :728

 ADD_POINTS_SUM_YR_1 ADD_POINTS_SUM_YR_2 EXCHANGE_COUNT     avg_discount
 Min.   :     0.0    Min.   :     0.0    Min.   : 0.0000   Min.   :0.0000
 1st Qu.:     0.0    1st Qu.:     0.0    1st Qu.: 0.0000   1st Qu.:0.6120
 Median :     0.0    Median :     0.0    Median : 0.0000   Median :0.7119
 Mean   :   540.3    Mean   :   814.7    Mean   : 0.3198   Mean   :0.7216
 3rd Qu.:     0.0    3rd Qu.:     0.0    3rd Qu.: 0.0000   3rd Qu.:0.8095
 Max.   :600000.0    Max.   :728282.0    Max.   :46.0000   Max.   :1.5000
```





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、平均折扣不为0、总飞行公里数大于0的记录。

```
# 数据清洗
# 丢弃票价为空的记录
delet_na <- datafile[-which(is.na(datafile$SUM_YR_1) | is.na(datafile$SUM_YR_2)),]
# 丢弃票价为0、平均折扣不为0、总飞行公里数大于0的记录
index <- ((delet_na$SUM_YR_1==0 & delet_na$SUM_YR_2==0)*
        (delet_na$avg_discount != 0)*
        (delet_na$SEG_KM_SUM > 0))
deletdata <-delet_na[-which(index==1),]
```

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。
经过属性选择后的数据集如下：

Show 25 entries Search:

	LOAD_TIME	FFP_DATE	LAST_TO_END	FLIGHT_COUNT	SEG_KM_SUM	avg_discount
1	2014/3/31	2006/11/2	1	210	580717	0.961639043
2	2014/3/31	2007/2/19	7	140	293678	1.25231444
3	2014/3/31	2007/2/1	11	135	283712	1.254675516
4	2014/3/31	2008/8/22	97	23	281336	1.090869565
5	2014/3/31	2009/4/10	5	152	309928	0.970657895
6	2014/3/31	2008/2/10	79	92	294585	0.967692483
7	2014/3/31	2006/3/22	1	101	287042	0.965346535
8	2014/3/31	2010/4/9	3	73	287230	0.962070222
9	2014/3/31	2011/6/7	6	56	321489	0.828478237
10	2014/3/31	2010/7/5	15	64	375074	0.708010153
11	2014/3/31	2010/11/18	22	43	262013	0.988658044
12	2014/3/31	2004/11/13	6	145	271438	0.95253487
13	2014/3/31	2006/11/23	67	29	321529	0.799126984
14	2014/3/31	2006/10/25	3	118	179514	1.398381742
15	2014/3/31	2010/2/1	2	50	270067	0.921984841
16	2014/3/31	2008/3/28	65	22	234721	1.026084586
17	2014/3/31	2010/7/15	7	101	172231	1.3865249
18	2014/3/31	2010/11/10	45	40	284160	0.837844243
19	2014/3/31	2006/4/6	2	64	169358	1.401596264
20	2014/3/31	2011/8/29	24	38	332896	0.70828541

Showing 1 to 25 of 62,051 entries Previous 1 2 3 4 5 … 2483 Next

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。
构建LRFMC5个指标：

L

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

R

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份=最后一次乘机时间至观测窗口末端时间长（单位：月）

F

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

M

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）

C

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）

标准化处理

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，发现指标取值范围数据差异较大，为了消除量纲带来的影响，需要对数据进行标准化处理。

Show 25 entries

Search:

ZL	ZR	ZF	ZM	ZC
1.69	0.14	-0.636	0.069	-0.337
1.69	-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
0.89	-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
-0.868	-0.268	-0.281	-0.262	3.396
-1.075	0.025	-0.423	-0.521	0.15
1.907	-0.884	2.979	2.13	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	-0.14	-0.22	-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	-0.5
-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
-0.15	-0.521	1.278	1.392	1.124

Showing 1 to 25 of 62,044 entries

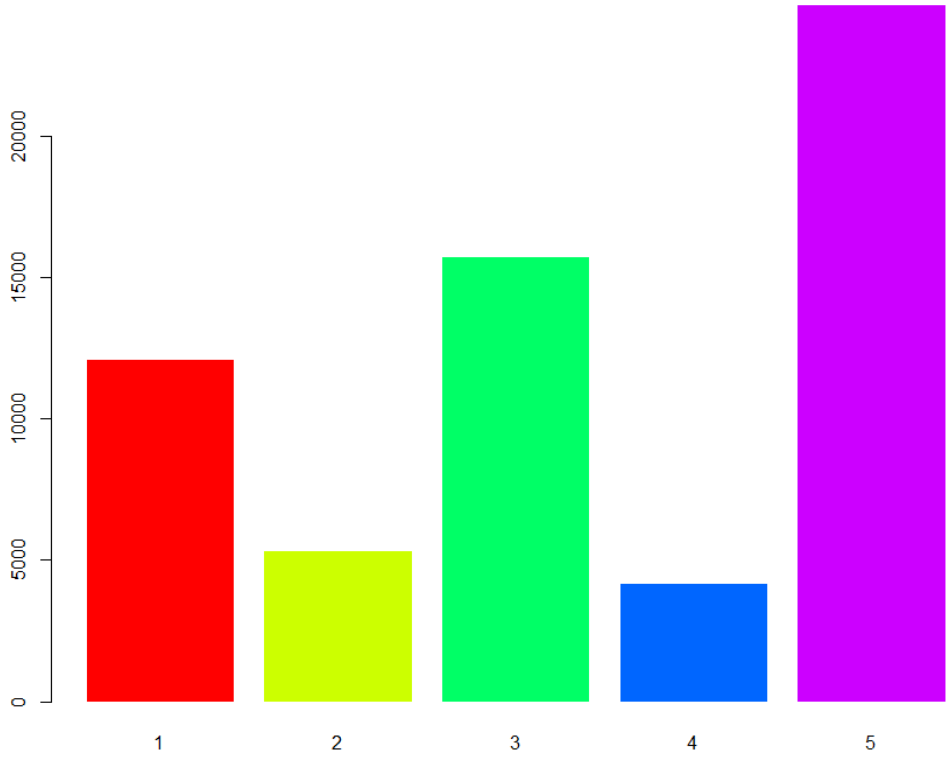
Previous 1 2 3 4 5 … 2482 Next

模型构建

客户价值分析模型构建主要由两部分构成，第一部分根据航空公司客户五个指标的数据，对客户做聚类分群；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。

```
# 分成五类
result <- kmeans(zscoreddata,5)
# 查看类别分布
table(result$cluster)
barplot(table(result$cluster),col= rainbow(5),border = F)
# 查看类中心
result$centers
```

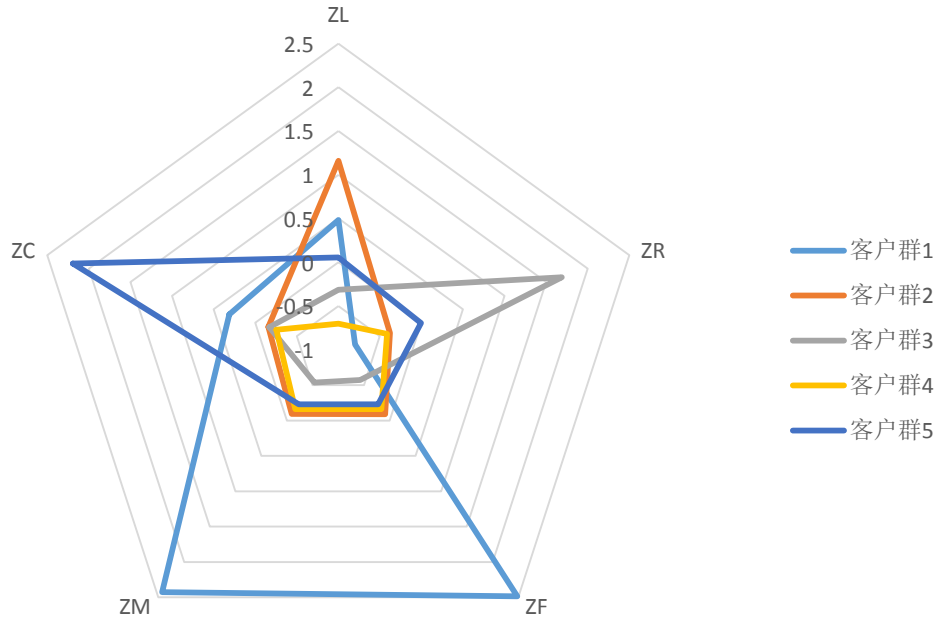
	ZL	ZR	ZF	ZM	ZC
1	-0.31409450	1.685902246	-0.57395741	-0.53677922	-0.1707754
2	0.48290520	-0.799404592	2.48309523	2.42434109	0.3077151
3	1.16037666	-0.377172347	-0.08690395	-0.09495047	-0.1579890
4	0.05681747	-0.005616371	-0.22678236	-0.23019920	2.1912013
5	-0.70054964	-0.414915903	-0.16112278	-0.16097852	-0.2549698



客户价值分析

针对聚类结果进行特征分析，绘制雷达图。

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	1.16	-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	-0.161	-0.255
客户群5	4198	0.057	-0.006	-0.227	-0.23	2.191



客户群1在F、M属性上最大，在R属性上最小，因此可以说F、M、R在客户群1上是优势特征；以此类推，F、M、R在客户群3上是劣势和弱势特征。

客户群特征描述表

根据雷达图总结出每个群的优势和劣势特征。

群别特征	优势特征			劣势特征		
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：正常字体标识最大值，加粗字体标识次大值，斜体字体标识最小值，带下划线的字体标识次小值

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。

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