**2016 数据挖掘Project**

**选题：航空公司客户价值分析**

**组号：第一题第21组**

**组员：李海锋 (201628013229015)**

**姜云涛 (201628013229013)**

**何平 (201628013229009)**

**梁胜文（201628013229019）**

**1．引言**

本次作业的主要目标是对航空公司的客户进行客户价值分析，建立相应的客户价值模型。根建立的客户价值模型，对航空公司的客户进行分类，根据商业上经典的五级客户分类模型，可以将客户分为潜在客户、目标客户、准客户、成交客户、忠诚客户。同时，根据美国数据库营销研究所Arthur Hughes 的研究，客户数据库中有三个神奇的要素，分别为最近一次消费（Recency）、消费频率（Frequency）、消费金额（Monetary）,这三个要素构成了经典的RFM模型，该模型是衡量客户价值和客户创利能力的重要工具与手段，在客户关系管理（CRM）中，占有重要的地位。

根据航空公司的数据属性值显示，数据集包含客户基本信息，乘机信息，积分信息等。在客户基本信息中，多数信息是无用的，例如会员卡号，性别，工作地城市等。依据RFM模型，我们基本可以排除客户基本信息中的全部信息。但是，考虑到客户成为航空公司会员时间的长短（入会时间到窗口观测时间）能够作为客户价值衡量的一种手段。因此，我们扩展了传统的RFM模型，增加了第四种衡量标准——客户关系长度（L）。同时，根据RFM模型，我们可以看出在乘机信息中，我们能够提取到观测窗口时间内的累积飞行里程（M），客户最近一次乘坐飞机距离观测窗口结束的月数（R），客户在观测窗口内乘坐公司飞机的次数（F）三个指标。在乘机信息中，平均折扣率这个属性，是反应客户价值的关键属性，当客户折扣率较高时，可能该客户的价值就偏低。因此，我们引入了AVG\_DISCOUNT(平均折扣率)作为模型建立的第五个衡量标准。在积分信息中，最重要的为总累计积分值，考虑到航空公司的积分值，一般与客户的累计飞行里程挂钩，因此这里则不引入该属性值作为本次数据模型建立的属性值。

因此，在这里我们在RFM模型的基础上，加入了入会时间长度（L），消费时间间隔（R），消费频率（F）、总飞行里程（M），折扣系数平均值（C）五种衡量指标。

**表 1.1 LRFMC模型**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 衡量指标 | L | R | F | M | C |
| 说明 | 会员入会时间到观测窗口的时间间隔（单位：月） | 客户最近一次乘坐公司飞机距离观测窗口结束月数的时间间隔（单位：月） | 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值 |

根据LRFMC模型，同时采用商业上经典的客户分类方法，将客户分为潜在客户、目标客户、准客户、成交客户、忠诚客户五种类型。根据已有分类数量的标准，分析各种分类算法的特点，我们采用了K-Means算法，这里，K=5,对应着我们对客户的五种分类模式。

根据观察航空公司提供的数据集，我们发现数据中存在属性为空值，属性残缺等问题。因此，本次数据挖掘的主要流程如图1.1所示。

缺失值与异常值分析

航空公司数据集

数据清洗

数据规约

数据标准化

K-means客户分类，模型建立

模型分析，客户价值分析排名

根据所得模型，提出相应的促销策略

**图1.1 航空客户价值分析总体流程**

第一步、如图1.1所示，因为航空公司数据集中，存在数据缺失，数据为空值等问题，因此首先对数据进行缺失值与异常值分析。

第二步、根据第一步的缺失值与异常值分析结果，我们进行相应的数据清洗，去除数据噪声，避免对模型的建立产生干扰。

第三步、清洗后的数据中包含了大量的冗余信息，根据我们采用的LRFMC模型，这里我们只需要数据集中的六种属性值，因此，这里我们通过数据规约，去除冗余值，得到相应的L、R、F、M、C的数值。

第四步、根据数据挖掘的基本步骤，同时避免数据值分散对模型建立的干扰，这里我们将进行数据的标准化，根据标准化的基本方法，这里我们采用了Z\_SCORE标准化方法。

第五步、将标准化好后数据进行K-means分类，得到相应的分类模型。

第六步、对模型进行分析，得出不同价值客户的属性特点，针对这些特点，提出相应的营销策略，制定相应的服务方案，提高航空公司的营业额。

下面将分布介绍每一步的具体实现流程与算法。

**2.1 数据缺失值与异常值分析**

由于数据存在大量的不规范和缺失，首先需要进行数据的缺失值和异常值的分析，标记出那些属性上存在缺失值与异常值，缺失数量，异常通过判断数值属性是否是数字字符串和其最大值与最小值是否超过定义域来判断，同时标记与规定属性数量不一致的数据行，为数据的清洗做必要的准备。

具体流程：读取数据文件的每一行，以','进行split，判断所有属性是否存在缺失，对于数值属性判断是否是数字，统计每个属性的缺失值数量和输指数型的最大值与最小值。

**2.2 数据预处理**

**2.2.1 数据清洗**

通过上面的数据缺失值与异常值分析，发现原始数据中存在一部分缺失值，同时还有着票价最小值最小为0，折扣率最小值为0，总飞行公里数大于零的记录。由于原始数据量大，这类数据所占比例较小，对该问题影响不大，因此对其进行丢弃处理，具体处理方法如下：

1、丢弃票价为空的记录。

2、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。

具体方法：对满足清洗条件的数据进行丢弃，处理方法为满足清洗条件的一行数据全部丢弃，记录需要丢弃的数据行数，将其置空。

**2.2.2 属性规约**

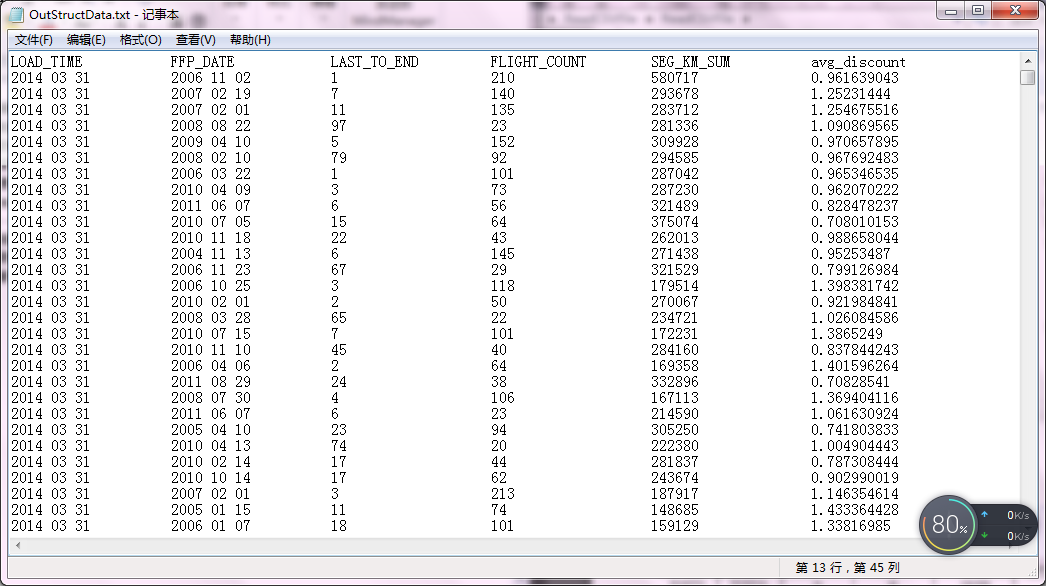
当数据清洗完成后，这里我们将进行相应的属性规约操作，根据LFRMC模型，在数据集中，我们需要提取出FFP\_DATE、LOAD\_TIME、FLIGHT\_COUNT、AVG\_DISCOUNT、SEG\_KM\_SUM、LAST\_TO\_END六种属性值，删除其余不相关的所有属性值。

算法实现：

根据数据集，结合CSV文件的特点，判断出相应属性值所在的位置；

根据属性值所在位置，读取数值，将其存放在一个LRFMC结构体数组中。

经过属性规约后的数据集如图1.2所示。



**图1.2 属性规约后的数据集**

**2.2.3 LRFMC值计算**

根据LRFMC模型，这里我们需要计算出L、R、F、M、C相应的属性值，根据各自的定义，我们可以得到：L = 观测结束时间到会员入会时间的间隔，R = 最后一次乘机时间到观测窗口结束的时间，F = 观测窗口内飞行的次数，M = 观测窗口内飞行的总公里数，C= 平均折扣率。

因此，根据属性规约的值，我们得到如下计算方法：

L = (LOAD\_TIME – FFP\_DATE) / 30;

R = LAST\_TO\_END/30;

F = FLIGHT\_COUNT;

M = SEG\_KM\_SUM;

C = AVG\_DISCOUNT;

上式中，L的计算值以月为单位，因此，首先计算间隔天数，然后除以30天，作为月数的间隔。R中同样除以30天作为月数的计算基本单位。

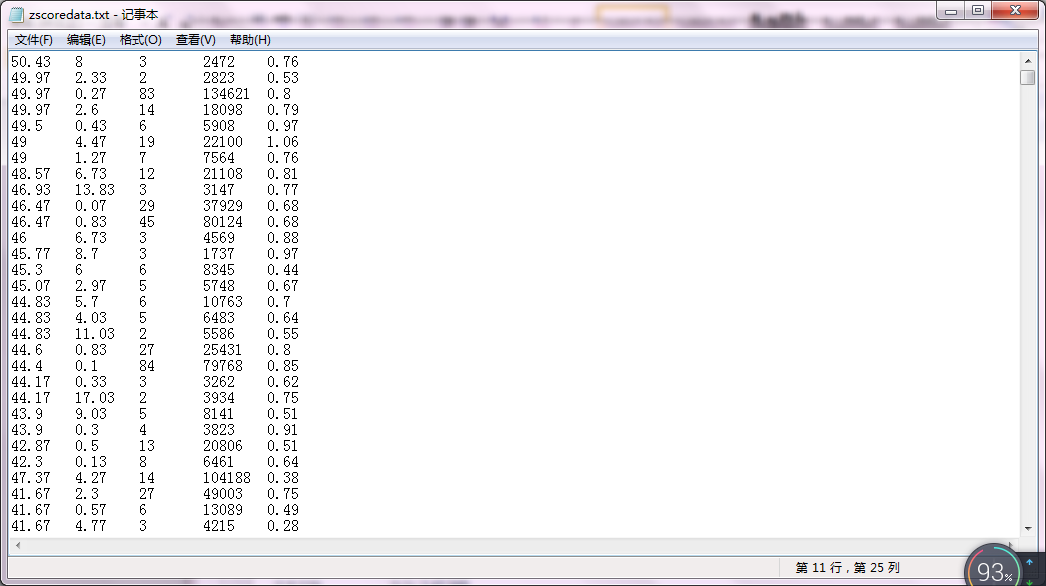
算法实现：

首先，将字符串2014/3/31,以/作为分隔符，分成年月日，存放在Date结构体数组中；

利用L = （LOAD\_TIME – FFP\_DATE）/30 ,首先计算出日期间隔天数，后面除以30，计算出相应的L值。

R、F、M、C通过相应的赋值语句得到。

计算后的数据集结果图1.3所示

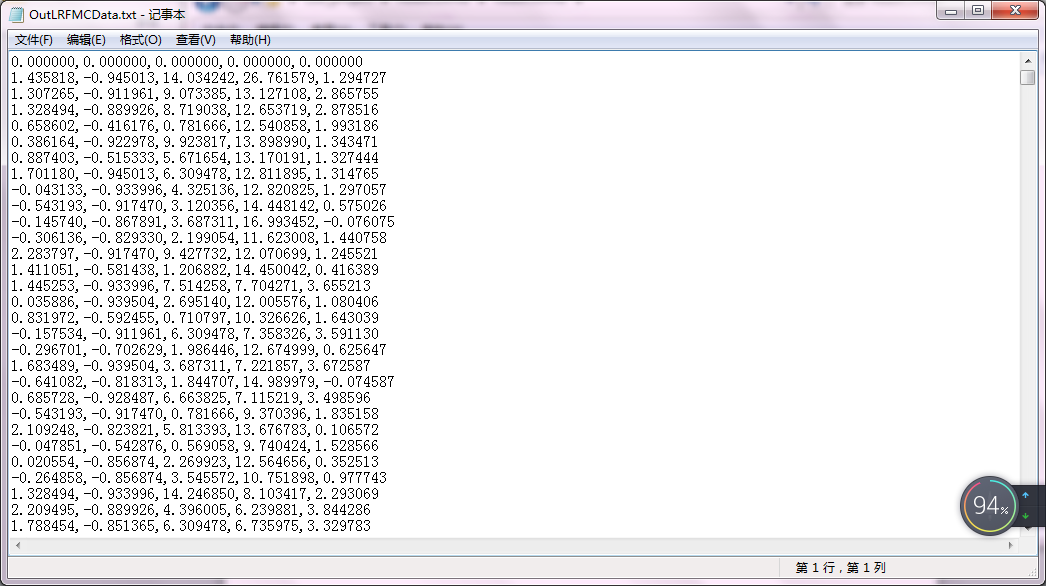


**图1.3 LRFMC数值计算结果**

**2.2.4 数值标准化**

这里从众多标准化的方法中，选择了Z\_score标准化。计算公式为：

根据上式，我们分别将L、R、F、M、C数值进行了标准化，标准化后的结果如图1.4所示。



**图1.4 数值标准化后结果**

**3.1 LRFMC客户模型生成**

根据上面提到的五种客户类型，这里我们采用了K-means算法，将K值赋值为5，然后进行客户分类。

k-means算法具体流程如下：

输入：k, data[n]

（1）选择k个初始中心点，例如c[0]=data[0],…c[k-1]=data[k-1]；

（2）对于data[0]….data[n]，分别与c[0]…c[k-1]比较，假定与c[i]差值最少，就标记为i；

（3）对于所有标记为i点，重新计算c[i]={ 所有标记为i的data[j]之和}/标记为i的个数；

（4）重复(2)(3)，直到所有c[i]值的变化小于给定[阈值](http://baike.baidu.com/view/409216.htm)。

由于K-means 算法的受到K值和初始点的选取，因此这里我们做出了改进，采用了K-means++算法，该算法通过选取距离尽量足够远的分类中心点，来确定初始的分类中心点，避免人为选取造成的干扰。

K\_means++算法的具体流程如下：

1、从输入的数据点集合中随机选择一个点作为第一个聚类中心

2、对于数据集中的每一个点x，计算它与最近聚类中心(指已选择的聚类中心)的距离D(x)

3、选择一个新的数据点作为新的聚类中心，选择的原则是：D(x)较大的点，被选取作为聚类中心的概率较大

4、重复2和3直到k个聚类中心被选出来

5、利用这k个初始的聚类中心来运行标准的k-means算法

**代码如下图：**

// use k-means ++ method

voidInitializeCorKMeans(){

for (inti = 0; i<sampleNum; i ++) {

sample[i].Clas = -1;

}

inttempIndex = (int)((sampleNum - 1) \* rand() / (RAND\_MAX+0.1)); //产生 0 - (sampleNum-1)的随机数，使其作为seed

for (inti = 0; i<dimension; i ++) {

coreNode[0].address[i] = sample[tempIndex].address[i];

}

doublelen[sampleNum];

memset(len, 0, sizeof(len));

//计算另外k-1个初始中心

for (inti = 1; i<k; i ++) {

doublesumLen = 0;

//更新每个sample距离D(x)

for (int j = 0; j <sampleNum; j ++) {

doubleminlen = 99999;

for (int m = 0; m <i; m ++) {

if(minlen>distance(coreNode[m], sample[j])){

minlen = distance(coreNode[m], sample[j]);

}

}

len[j] = minlen;

sumLen += len[j];

}

double random = (sumLen) \* rand() / (RAND\_MAX+0.1); // 产生0-sum(D(x))的随机值random

for (int j = 0; j <sampleNum; j ++) {

if(random <0){

for (int m = 0; m <dimension; m ++) {

coreNode[i].address[m] = sample[j].address[m];

}

break;

}

random -= len[j];

}

}

}

用K-means++算法进行分类后，结果如下表：

表 1.2 算法输出结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **客户** | **类别** | **中心点** | | | | |
| **类别** | **个数** | **L** | **R** | **M** | **F** | **C** |
| **类别1** | **15807** | **1.158505** | **-0.375845** | **-0.089081** | **-0.096903** | **-0.152706** |
| **类别2** | **5325** | **0.482255** | **-0.799382** | **2.485566** | **2.425889** | **0.302865** |
| **类别3** | **12195** | **-0.314916** | **1.682692** | **-0.573924** | **-0.537001** | **-0.166581** |
| **类别4** | **24899** | **-0.699468** | **-0.412737** | **-0.163175** | **-0.163749** | **-0.239461** |
| **类别5** | **3822** | **0.098363** | **-0.012033** | **-0.200326** | **-0.198902** | **2.301111** |

**3.2 客户价值分析**

由表1.2所示，图中显示了每个群体的聚类数目，同时记录了相应类别的聚类中心。通过对每个类别的中心值进行对比。

根据航空公司实际情况与LFRMC模型，我们可以确定，飞行次数（F）、总飞行里程（M）、会员入会时间（L）的值越大，则该客户则较重要，同时，最近乘机时间间隔也是衡量的重要指标，该指标可以帮助航空公司避免流失客户。

由图1.5与表1.6，我们可以发现客户群体1的L属性值最大，客户群体2的F、M属性值最大，R属性值最小。客户群体3的R属性值最大，但是F、M属性值最小，客户群体4的L、C属性值最小，客户群体5的C属性值最大。

**图 1.5五类客户折线图**

**图 1.6属性对应折线图**

**图 1.7五类客户雷达图**

由图1.7可知，对于第一类客户，L、F、M的值较大，对于该客户群体则为优势特征，

对于第二类客户，F、M的值较大，同时R值很小，则可以确定F、M、R为第二类客户的优势特征。那么对于客户群体3,4而言，则出现相对的弱势特征，对于客户群体3，我们可以看到F、M、R为该群体的弱势特征，对于客户群体4，L、C为该客户群体的弱势特征。那么对于客户群体5而言，由于R值较高，为该群体的优势特征，但侧面反映该客户的重要程度相比于客户群体1,2,3,4而言较低。

根据表1.2所示可以看出，数据中包含负数，这里，为了便于数据对比，我们将对比标准从Y =0 ,偏移到Y = 1；由此，可以推出相应的图表数据为：

**图1.8 五种客户类型柱状图**

根据商业上经典的五种客户类型：潜在客户、目标客户、准客户、成交客户、忠诚客户，我们可以根据上述的五张图表来分析出每一种类别客户所属的客户类型。

对于第一类客户而言，该客户的L值较高，因此但是F、M、C值相对较高，因此可以归纳为准客户，此种类客户需要采取措施给予挽留，避免客户流失向其他航空公司。

对于第二类客户，该种客户的F、M、C很高，并且R值很低，为忠诚客户，是航空公司需要重点保持的客户，该种客户能够为公司带来长期利益，因此需要航空公司采取更加优秀的服务来让此类客户享受到更加舒适的服务。

对于第三种客户而言，他们R值相对最高，并且其他属性值偏低，说明这些客户对航空公司而言，价值不是很高，但是可能是相应的潜在客户，给予一定的优惠政策，可能从潜在客户发展成为目标客户。

对于第四类客户，他们的属性值没有什么特别突出的特点，因此可以归纳为一般客户，对航空公司而言，采取一定的优惠政策，可以使得该种类客户发展成为目标客户，因此归纳为目标客户。

对于第五类客户而言，他们的C值很高，并且其他属性值相对较高，因此可以归纳为成交客户，该种类客户偏向于廉价的航空旅行，因此，航空公司采取一定的优惠政策，便可以让该种类客户购买航空公司的机票，因此是需要重点发展的客户。

根据以上五种客户类型，我们归纳出了表1.3

**表1.3客户类型分布**

|  |  |
| --- | --- |
| 客户所述类别 | 客户类型 |
| 第二类客户 | 忠诚客户 |
| 第五类客户 | 成交客户 |
| 第一类客户 | 准客户 |
| 第四类客户 | 目标客户 |
| 第三类客户 | 潜在客户 |

**3.3营销策略**

1、对于忠诚客户而言，该类客户的特点在于，乘坐飞机的次数很多，并且不会考虑飞机价格的高低。他们看中的是航空公司的服务，与航空公司航班设置时间安排的好坏。因此对于此类客户，航空公司应该定期推出针对该种客户的优惠政策，例如赠送落地酒店住宿券、家人乘坐航班优惠券等较为高级的优惠政策。同时，在该种类客户乘坐飞机时，应该着重考虑对该种类的客户的服务，进行针对化服务，例如，专用候机室，专用乘机通道等，提高此种类型客户对航空公司的好感度与粘着性。

2、对于成交客户与准客户而言，此种类型的客户，可能近期乘坐公司航班的可能性降低，此时，提高对该种类型客户的优惠额度，可能会提高该种类客户乘坐飞机的频率。

3、对于目标客户与潜在客户而言，他们乘坐飞机的次数较低，并且对于飞机票价格具有相对较高的敏感性。可能大多是旅游类型的客户，因此针对该种类客户，航空公司最主要的是在旅游旺季，为该种类客户进行一定的折扣，提高该种类客户的对该航空公司的黏着性。同时，推出相应的旅游优惠方案，为客户提供一条龙服务。

根据以上分析，结合航空公司特点，具体营销策略如下：

1、实行航空公司会员制度，采取会员积分制度，不同等级的客户，拥有不同等级的会员卡。对于忠诚客户，可能享受航空公司会员的最高待遇。比如，专用候机楼等。

2、在旅游旺季时候，推出一定量的旅游特价机票方案，可以与目的地酒店结合，制定相应的酒店机票一体化套餐，为旅游人士提供最好的服务。

3、在旅游淡季时候，推出特价机票，并在城市的商圈进行推广，同时在各大高校推广，为未来的乘机客户打下基础，提高后续航空公司的影响力。