****

**《金融数据挖掘案例分析》**

**课程设计报告**

题 目： 基于R语言的公司客户关系管理

学生姓名： xxx 序号： 01

学生姓名： xxx 序号： 01

学生姓名： xxx 序号： 01

学生姓名： xxx 序号： 01

指导教师：  许美玲

信息管理与工程系

2020年06月3日

**基于R语言的公司客户关系管理**

**摘 要：**

**关键词：**R；决策树算法；风险预测

**Based on R Language in Customer Relationship Management of Airlines**

**Abstract:**

**Key words:**

**目录**

[1数据挖掘背景及目标 1](#_Toc42266732)

[1.1背景： 1](#_Toc42266733)

[1.2挖掘目标 2](#_Toc42266734)

[2数据理解 3](#_Toc42266735)

[3数据预处理 5](#_Toc42266736)

[4模型的建立 11](#_Toc42266737)

[4.1 Kmeans聚类模型 11](#_Toc42266738)

[4.2NaiveBayes分类模型 17](#_Toc42266739)

[5模型评估及改进 20](#_Toc42266740)

[5.1模型评估 20](#_Toc42266741)

[5.2模型改进 22](#_Toc42266742)

[6结论 23](#_Toc42266743)

[7系统收获 27](#_Toc42266744)

[8参考文献 28](#_Toc42266745)

[附录 29](#_Toc42266746)

# 1数据挖掘背景及目标

## 1.1背景：

1）行业内竞争

民航的竞争除了三大航空公司之间的竞争外，还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。

IMG_256IMG_256

图1.各国主要的民航公司

2）行业外竞争

随着高铁、动车等铁路运输的兴建，航空公司受到巨大的冲击。

3）了解客户价值分析

为了区分无价值客户、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，并制定相应的营销策略，将有限的营销资源集中于高价值客户，对不同的客户提供个性化的客户服务，实现企业的利润最大化目标。

了解客户价值分析：

客户营销战略倡导者Jay&AdamCurry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。

如何分析客户价值：

RFM模型是衡量客户价值和客户创利能力的重要工具和手段。在众多的客户关系管理(CRM)的分析模式中，RFM模型是被广泛提到的。该机械模型通过一个客户的近期购买行为、购买的总体频率以及花了多少钱3项指标（最近消费时间间隔（Recency）、消费频率（Frequency）和消费金额（Monetary））来描述该客户的价值状况，从而识别出高价值的客户；但是RFM模型并不适用所有的场景：比如航空票价受到运输距离、舱位等级等多种因素影响，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、高等级舱的旅客相比，后者对于航空公司的价值可能更高，因此若用RFM模型去分析航空公司客户，得出的结果一定是不准确的。  
 根据航空公司的业务设计对应的分析模型：  
 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对应的折扣系数的平均值C两个指标代替消费金额，考虑航空公司会员入会时间的长短在一定长度上影响客户价值，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。

## 1.2挖掘目标

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）对不同客户进行特征分析，比较不同类别客户的客户价值

3）建立客户分类模型，用来判断和预测客户的价值

4）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数据理解

航空公司客户数据说明：目前航空公司已经积累了大量的会员档案信息和其乘坐航班记录。以2018-3-31为结束时间，选取宽度为两年的时间段作为分析观测窗口，抽取观测窗口内有成绩记录的所有客户的详细数据行程里历史数据，44个特征，总共62988记录。数据特征及其说明如下表所示：

表1.航空公司客户数据说明：

|  |  |  |
| --- | --- | --- |
| **表名** | **特征名称** | **特征说明** |
| **客户基本信息** | 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 |
| **乘机信息** | 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 | |
| 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 | |
| LAST\_TO\_END | 最后一次乘机时间至观测窗口结束时间长 | |
| 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 | |
| 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 | |
| PROMOPTIVE\_SUM | 促销积分 | |
| PARTNER\_SUM | 合作伙伴积分 | |

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记做LRFMC模型。  
航空公司LRFMC模型:

L："LOAD\_TIME"-"FFP\_DATE"会员入会时间距观测窗口结束时间的月数（单位：月）

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（单位：月）  
 F："FLIGHT\_COUNT"客户在观测窗口内乘坐公司飞机的次数（单位：次）  
 M："SEG\_KM\_SUM"客户在观测窗口内累计的飞行里程（单位：公里）  
 C:"avg\_discount"客户在观测窗口内乘坐舱位所对应的折扣系数的平均值（平均折扣率）  
 本文采取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，从而识别出最有的价值的客户。然后使用NaiveBayes对客户分类。

航空公司客户价值分析的步骤：

1. 从航空公司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中的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特征;  
   4）对客户进行分类，针对模型结果得到不同价值的客户，采取不同的营销手段，提供定制化的服务。

# 3数据预处理

#读取数据

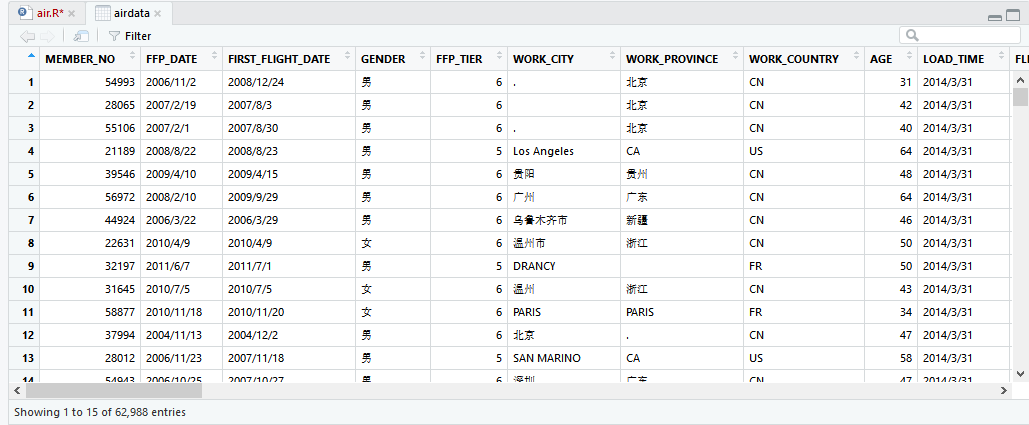
airdata<-read.csv("data//air\_data.csv",header=TRUE,sep=",")

图3.读取部分数据

#查看变量

names(airdata)

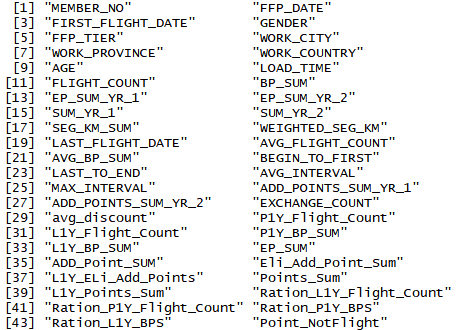


图4.变量名

##选择需要的变量

由于变量个数很多，许多都是分析中不需要的，可以选择性的对可能需要的变量进行探索性分析。

airdatanew<-airdata[,c(1:2,10:11,15:18,20:29)]

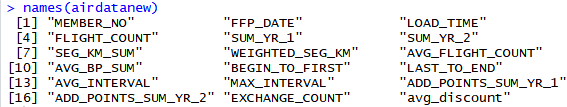


图5.选取重要数据

由于本分析中不涉及客户特征相关的分析，所以可以剔除。另外积分相关的数据也不涉及可以剔除。

#数据探索

summary(airdatanew)

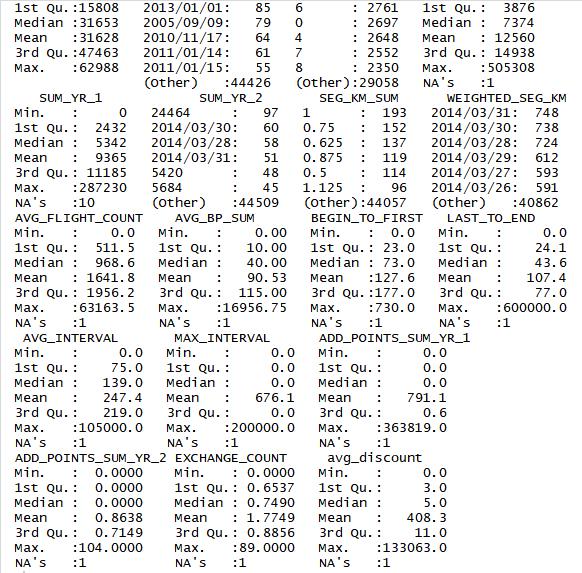


图6.查看数据分布情况

说明：

1、只有SUM\_YR\_1和SUM\_YR\_2存在缺失值

2、删除票价为0，但是平均折扣和总飞行公里数大于0的记录（逻辑错误值）

str(airdatanew)

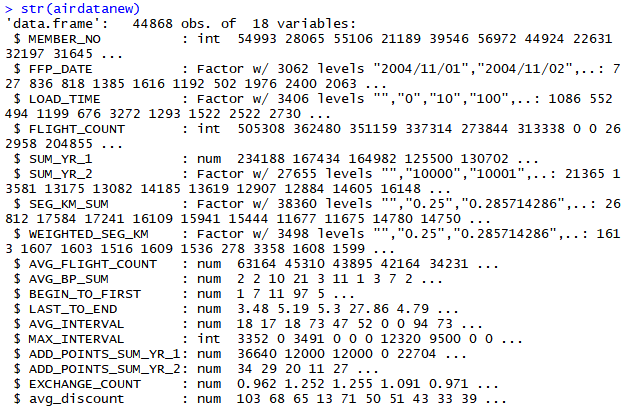


图7.部分新数据的变量情况

可以看出全部是连续型变量，FFP\_DATE和LOAD\_TIME都是因子，后面构建模型前需要将其转化为date格式，计算日期差。

#查看缺失值

install.packages("mice",dependencies=TRUE)

install.packages("VIM",dependencies=TRUE)

library(mice)

library(VIM)

md.pattern(airdatanew)

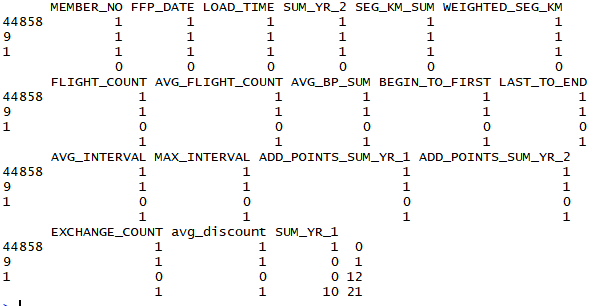


图8.数据空值情况

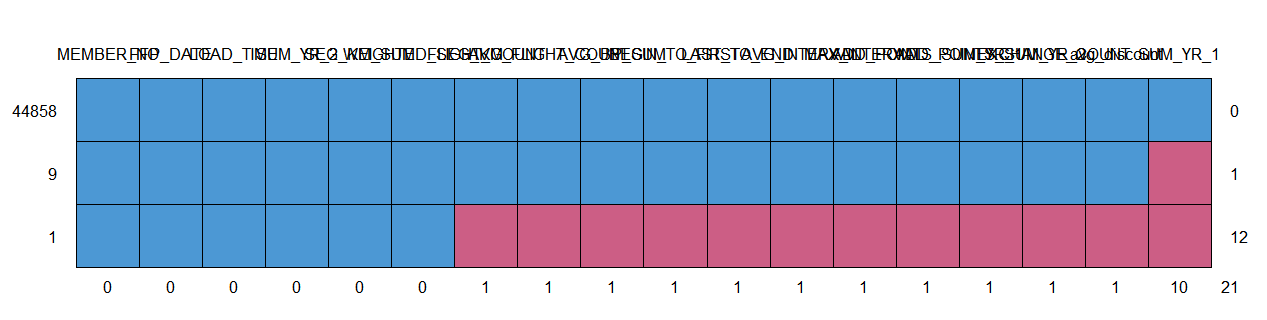


图9.数据空值情况

aggr(airdatanew,prop=F,numbers=T)

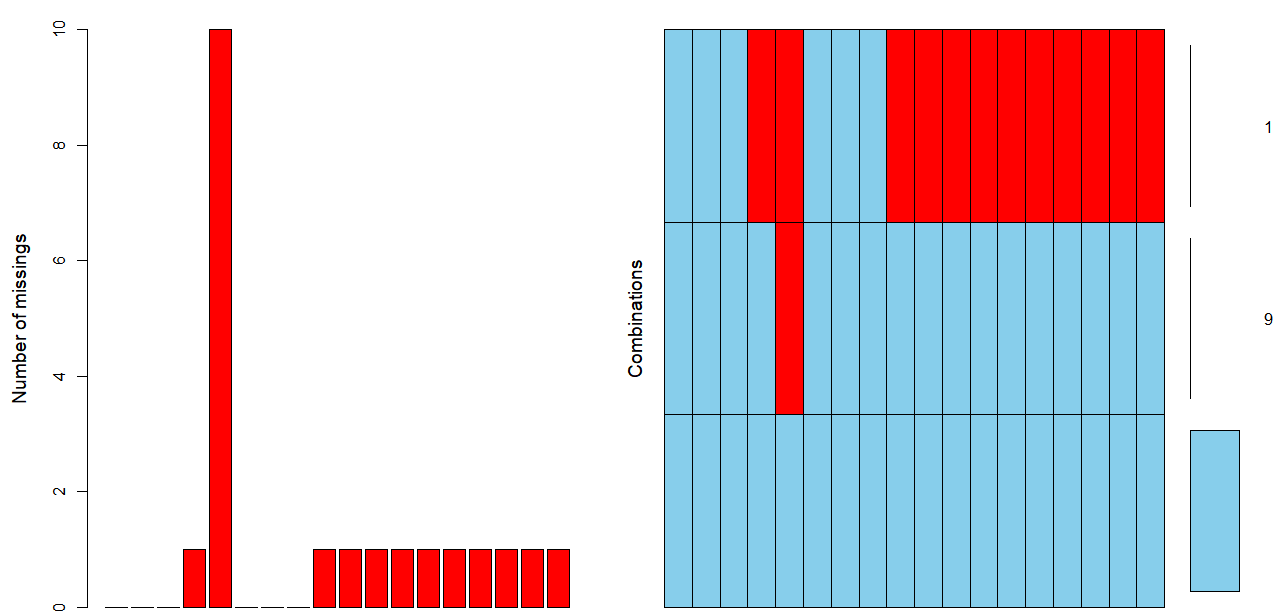


图10.数据空值情况

说明：只有两个变量有缺失值，且缺失值个数为138+551，较之于整体观测记录非常小可以直接删除。

#缺失值和不符逻辑值处理

#删除缺失值

airdatanew<-airdatanew[complete.cases(airdatanew),]#保留完整观测行

#去除逻辑值异常的观测，即删除票价为0，但是平均折扣和总飞行公里数大于0的记录

index<-which(airdatanew$SUM\_YR\_1==0&airdatanew$SUM\_YR\_2==0&airdatanew$avg\_discount!=0&airdatanew$SEG\_KM\_SUM>0)

airdatanew<-airdatanew[-index,]

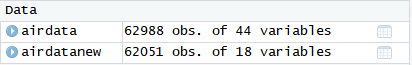


图11.新数据框

说明：删除缺失值和逻辑错误值后，剩下62051个观测。

##属性规约：根据LRFMC模型挑选有用的变量

airdatanew<-airdatanew[,c("MEMBER\_NO","FFP\_DATE","LOAD\_TIME","LAST\_TO\_END","FLIGHT\_COUNT","SEG\_KM\_SUM","avg\_discount")]

DT::datatable(airdatanew)（.表格生成函数）

说明：至此，分析所需变量以筛选和清洗出来，但是挑选出来的变量要带入模型需要进行变量转换。

##数据变换

#L="LOAD\_TIME"-"FFP\_DATE"(观测窗口的结束时间-入会时间)（单位：月）

#R="LAST\_TO\_END"（最后一次乘机时间至观测窗口末端时间长）（单位：月）

#F="FLIGHT\_COUNT"（观测窗口内的飞行次数）（单位：次）

#M="SEG\_KM\_SUM"（观测窗口的总飞行公里数）（单位：公里）

#C="avg\_discount"（平均折扣率）

#由于是因子变量，需要进行日期转换

airdatanew$FFP\_DATE<-as.Date(airdatanew$FFP\_DATE)

airdatanew$LOAD\_TIME<-as.Date(airdatanew$LOAD\_TIME)

#两日期相减得出的是天数，因而需要转化为月份数，这种转换可能和实际月份数有差别，但是不大。

airdatanew$L<-as.numeric(round((airdatanew$LOAD\_TIME-airdatanew$FFP\_DATE)/30,3))

#更改变量名

names(airdatanew)[4:7]<-c("R","F","M","C")

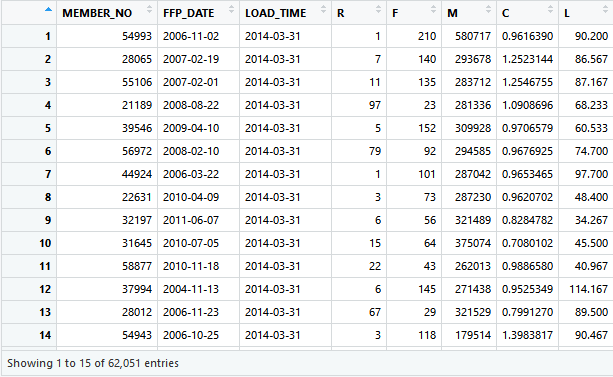


图12.新数据框

#将数据标准化处理

#除"MEMBER\_NO","FFP\_DATE","LOAD\_TIME"这3个字段，剩余字段构建模型

airdata\_scale<-scale(airdatanew[,-c(1:3)])

summary(airdata\_scale)

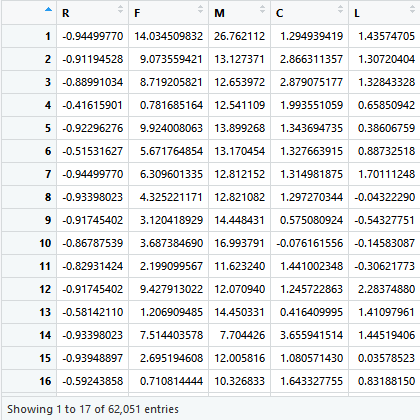
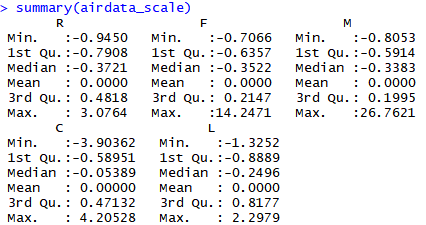


图13.将数据标准化处理

#将数据进行归一化处理，数据都在0-1之间分布

install.packages("caret",dependencies=TRUE)

library("caret")

#range函数的值是一个长度为2的向量，即c(min(x),max(x))

airdata\_range<-preProcess(airdatanew[,-c(1:3)],method=c("range"))

airdata\_range\_pre<-predict(airdata\_range,airdatanew[,-c(1:3)])

#将数据进行中心化处理

#数据中心化是数据标准化的步骤之一，即x-均值，其并没有消除量纲的影响。airdata\_center<-preProcess(airdatanew[,-c(1:3)],method=c("center"))

airdata\_center\_pre<-predict(airdata\_center,airdatanew[,-c(1:3)])

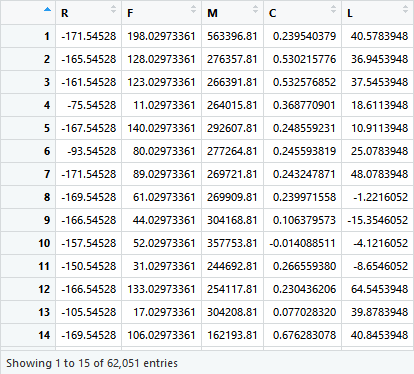
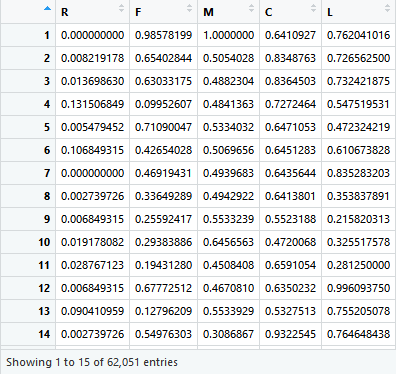


图14.将数据进行归一化和中心化处理结果

# 4模型的建立

## 4.1 Kmeans聚类模型

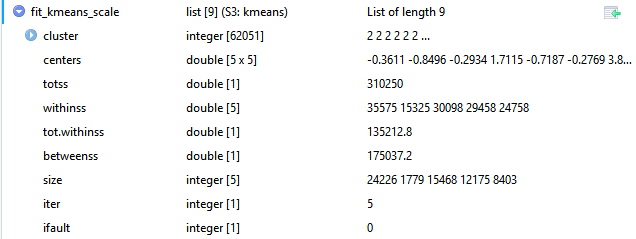
fit\_kmeans\_scale<-kmeans(airdata\_scale,5)

图15.标准化处理数据后的Kmeans聚类模型

fit\_kmeans\_range<-kmeans(airdata\_range\_pre,5)

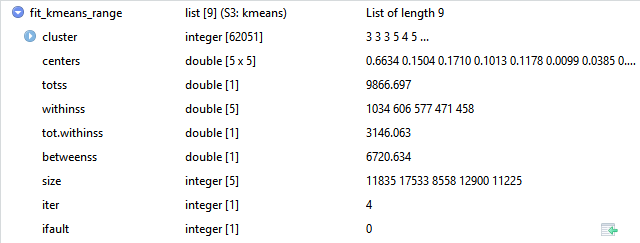


图16.归一化处理数据后的Kmeans聚类模型

fit\_kmeans\_center<-kmeans(airdata\_center\_pre,5)

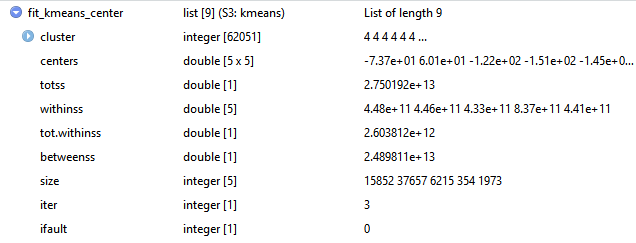


图15.中心化处理数据后的Kmeans聚类模型

#得出每个类别的中心

centers\_scale<-as.data.frame(fit\_kmeans\_scale$centers)#标准化处理

centers\_range<-as.data.frame(fit\_kmeans\_range$centers)#归一化处理

centers\_center<-as.data.frame(fit\_kmeans\_center$centers)#中心化处理

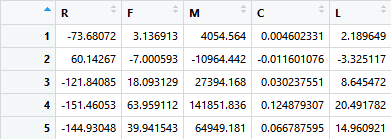
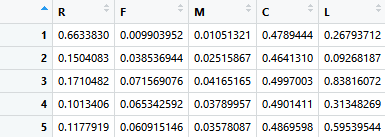
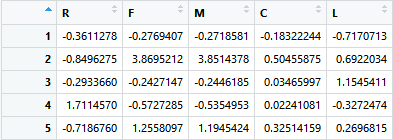


图16.标准化处理数据后的中心 图17.归一化处理数据后的中心 图18.中心化处理数据后的中心

说明：

归一化处理后，虽然数据都在0-1之间分布，但是从结果可以看出不同类别之间的区分度降低，特征反而不明显了；中心化处理后并没有消除量纲的影响，效果比较差，接下来不用考虑。

##存储每个类别的中心及频数

#标准化数据聚类结果

centers\_scale$freq<-as.data.frame(table(fit\_kmeans\_scale$cluster))[,2]



图19.标准化数据聚类结果

centers\_scale<-centers\_scale[,c("freq","L","R","F","M","C")]

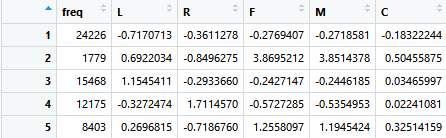


图20.标准化数据聚类结果

write.csv(centers\_scale,"centers\_scale.csv",row.names=T)#导出

##对该模型进行做图分析

#对freq列做条形图

barplot(centers\_scale$freq)

#使用ggradar绘制雷达图  
ggradar(centers\_scale)

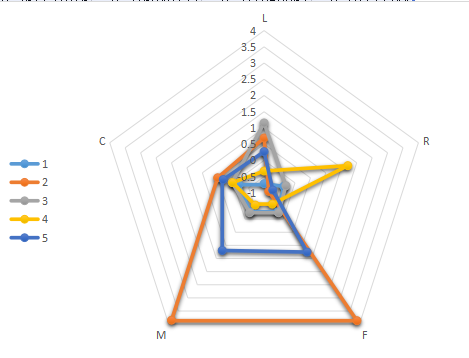
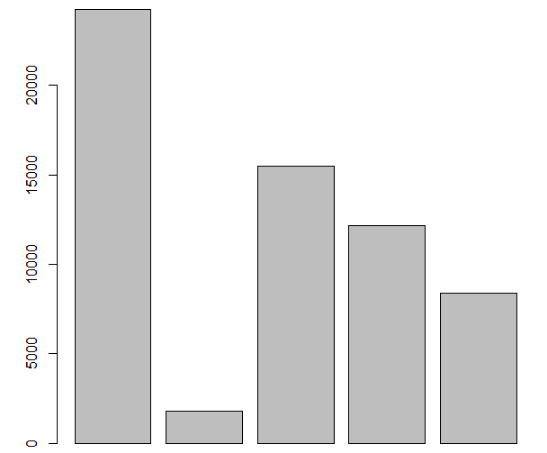


图21.对中心化数据freq列做条形图 图22.使用ggradar绘制中心化数据的雷达图

#归一化数据聚类结果

centers\_range$freq<-as.data.frame(table(fit\_kmeans\_range$cluster))[,2]

centers\_range<-centers\_range[,c("freq","L","R","F","M","C")]

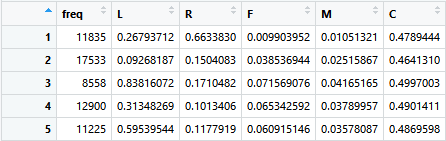


图23.归一化数据聚类结果

write.csv(centers\_range,"centers\_range.csv",row.names=T)

#作图分析

barplot(centers\_range$freq)

ggradar(centers\_range)

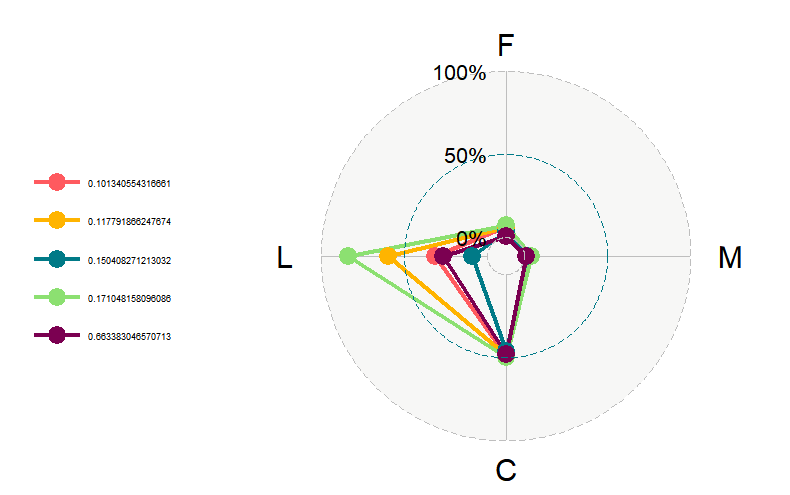
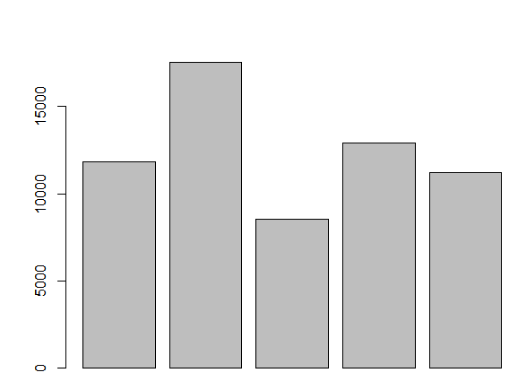


图24.对归一化数据freq列做条形图 图25.归一化数据绘制雷达图

说明：

通过两种数据预处理后的聚类分析结果对比,标准化后的数据聚类在LRFMC 5个变量上的分布更加明显，而归一化后的数据聚类结果主要差别集中在L和R上，所以分辨的效果不太好，所以采用标准化预处理效果比较好。

#如果需要对客户贴上聚类类别标签，则将分类标签组合到包含客户会员号的数据中

airdatanew$cluster<-as.data.frame(fit\_kmeans\_scale$cluster)

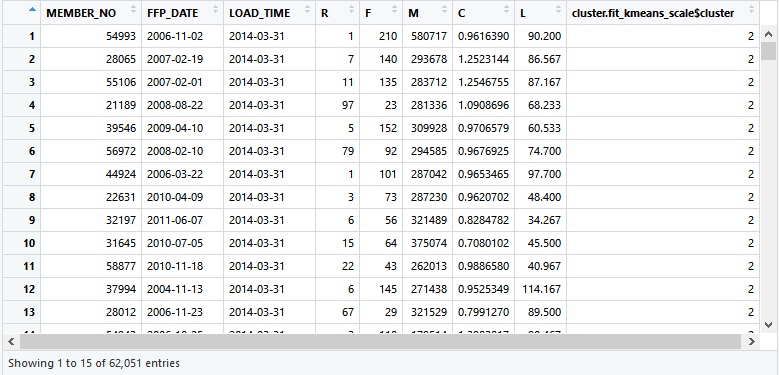


图26. 将分类标签组合到包含客户会员号的数据中

##聚类结果解读和应用

#L="LOAD\_TIME"-"FFP\_DATE"(观测窗口的结束时间-入会时间)（单位：月）

#R="LAST\_TO\_END"（最后一次乘机时间至观测窗口末端时间长）（单位：月）

#F="FLIGHT\_COUNT"（观测窗口内的飞行次数）（单位：次）

#M="SEG\_KM\_SUM"（观测窗口的总飞行公里数）（单位：公里）

#C="avg\_discount"（平均折扣率）

L和R越大，表示时间越长。L短表示新客户，意味着要激励，长意味着要挽留，在优劣势分析时，将L越长的设置为优势；C越大，表示折扣越低，那么暗示客户对价格不敏感；M和F越大，表示客户乘机需求越大。

⑴客户聚类结果

对数据进行聚类分群如下表所示：

表2：客户聚类结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 聚类类别 | 聚类  个数 | 聚类中心 | | | | |
| L | R | F | M | C |
| 群1 | 5336 | 0.48296123 | -0.79944875 | 2.48351216 | 2.42472275 | 0.3087468 |
| 群2 | 15745 | 1.16049784 | -0.37732916 | -0.37732916 | -0.09498699 | -0.1557095 |
| 群3 | 12134 | -0.31370040 | 1.68554920 | -0.57392324 | -0.53674300 | -0.1743790 |
| 群4 | 24666 | -0.70030743 | -0.41494899 | -0.16116894 | -0.16107783 | -0.2530821 |
| 群5 | 4170 | 0.05542063 | -0.00249783 | -0.22544385 | -0.22944009 | 2.1972694 |

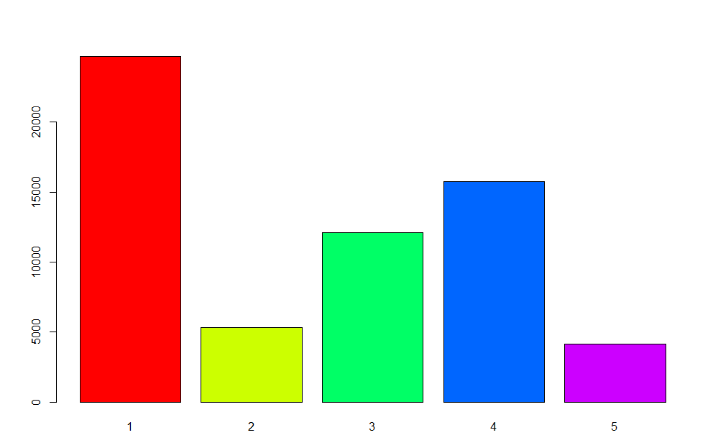
表3. 客户特征

|  |  |  |
| --- | --- | --- |
|  | **优势** | **劣势** |
| 1 | R | L、C |
| 2 | C | M、F、R |
| 3 | M、F、L |  |
| 4 |  | R、M、F |
| 5 | C、M、F、R |  |

 由上述的特征分析的图表说明每个客户群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。他们之间的区别如上图所示，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4：客户群分类

|  |  |  |
| --- | --- | --- |
| 客户群 | 排名 | 排名含义 |
| 客户群1 | 1 | 重要保持客户 |
| 客户群2 | 2 | 重要发展客户 |
| 客户群3 | 3 | 重要挽留客户 |
| 客户群4 | 4 | 一般客户 |
| 客户群5 | 5 | 低价值客户 |



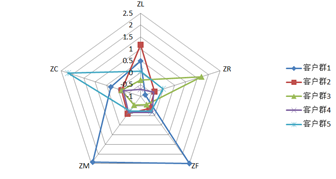


图27.客户分类的雷达图 图28：客户群特征分析图

## 4.2NaiveBayes分类模型

上面的KMeans聚类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。针对此问题，本文接下来采用NaiveBayes分类方法对客户类型进行过预测。

贝叶斯分类算法：

朴素贝叶斯分类法是一种生成学习算法。

假设：在y给定的条件下，各特征Xi 之间是相互独立的，即满足：

 (该算法朴素的体现之处)

原理： 贝叶斯公式

 （1）

思想：对于待分类样本，求出在该样本的各特征出现的条件下，其属于每种类别的概率（P(Yi|X)），哪种类别的概率大就将该样本判别为哪一种类别。

 （2）

 (P(X)为x的先验概率，与Y无关，在具体计算是分母可以直接忽略，只计算分子）

的估计:

（1）特征属性为离散值

    直接用每一类别中各名录出现的频率作为其概率值

（2）特征属性为连续性值

假设特征属性服从正太分布，用各类别的样本均值及标准差作为正态分布的参数。

 （3）

   （4）

airdatanew2<-airdatanew[,4:9]

ee<-as.matrix(airdatanew2$cluster)

airdatanew2$cluster<-as.vector(ee)

airdatanew2$cluster<-as.factor(airdatanew2$cluster)

说明：因为分类变量那一列类型是数据框，所以应当先转换为因子类型

##对数据进行分层，分为训练集和测试集

#使用随机抽样函数，取数据中的80%

inTrain=createDataPartition(airdatanew2$cluster,p=0.8,list=FALSE)

air\_Train=airdatanew2[inTrain,-6]

air\_Test=airdatanew2[-inTrain,-6]

air\_train\_labels=airdatanew2[inTrain,6]#训练集的标签

air\_test\_labels=airdatanew2[-inTrain,6]#测试集的标签

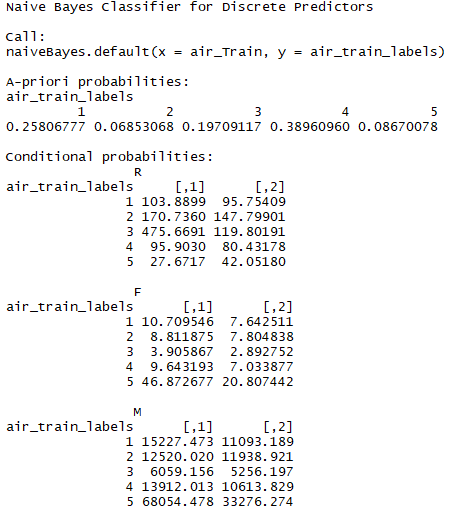
#建立NaiveBayes模型

install.packages("e1071",dependencies=TRUE)

library("e1071")

#使用训练集创建分类器

classifier=naiveBayes(air\_Train,air\_train\_labels)



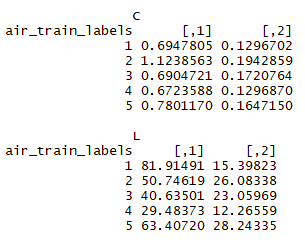


图29.NaiveBayes分类模型

#对测试集进行预测

p<-predict(classifier,air\_Test,type="class")

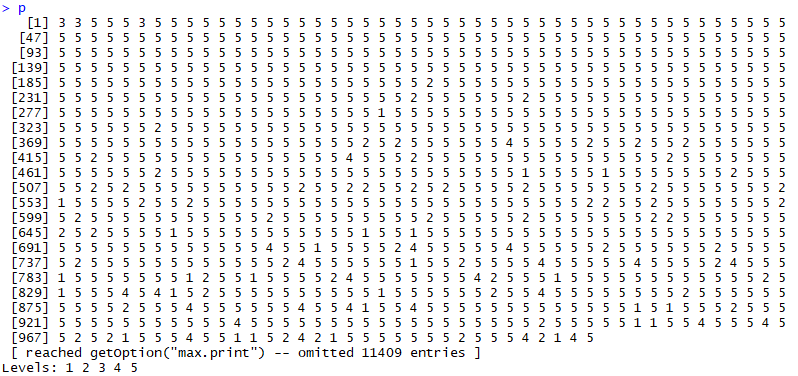


图30.测试集的预测结果

# 5模型评估及改进

## 5.1模型评估

1. 对贝叶斯模型检验使用table和混淆矩阵

#检验模型

t<-table(p,air\_test\_labels)

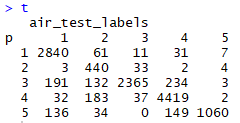


图31.测试结果与真实结果的对比

说明：

主对角线为正确值，其他的是预测错的值，可以看出来准确率还是比较高的，为了得到一个更加准确的正确率采用如下方式：

#正确率

paste0(round((sum(diag(t)))/sum(t),2)\*100,"%")

说明：正确率为90%

#构建混淆矩阵，评估模型的性能

library(gmodels)

CrossTable(p,air\_test\_labels)

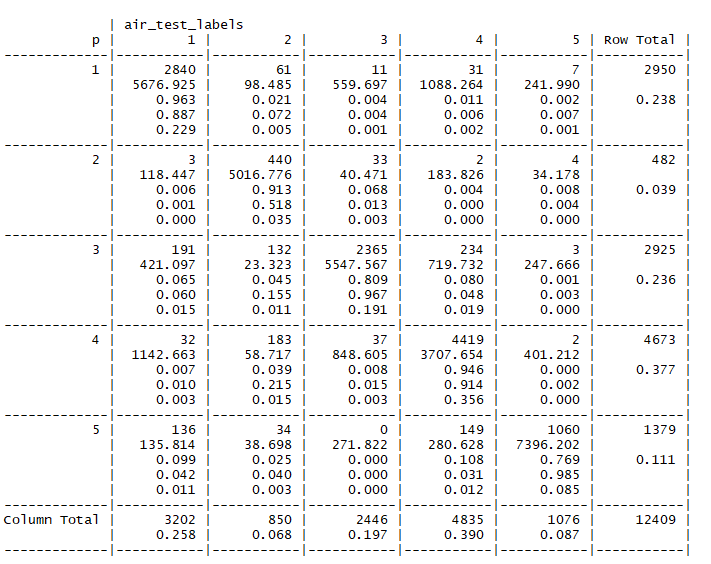
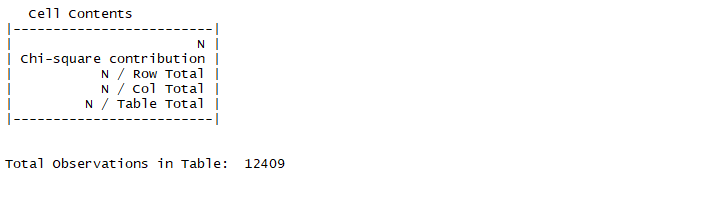


图32.混淆矩阵

说明：

混淆矩阵中：#第一个参数为预测值（x轴），第二个参数为真实值（y轴），构成矩阵的形式进行对比，主对角线为正确值的数量一共11124个正确分类的，一共有12409条数据以及其所对应的概率。

此次模型的正确率为90%，符合我们的建模要求，并且具有较高的正确率，说明模型具有很好的预测功能。

2）模型优点

Kmeans聚类优点：

a）是解决聚类问题的一种经典算法，简单、快速

b）对处理大数据集，该算法保持可伸缩性和高效性

c）当簇接近高斯分布时，它的效果较好。

NaiveBayes分类优点：

a)在接受大数据量训练和查询时所具备的的高速度。

b)对分类器实际学习状况的解释相对简单。

3）模型缺点

Kmeans聚类缺点：

a）在簇的平均值可被定义的情况下才能使用，可能不适用于某些应用；

b）在 K-means 算法中 K 事先给定，难以估计。很多时候，事先并不知道给定的数据集应该分成多少个类别才最合适；

c）在 K-means 算法中，首先需要根据初始聚类中心来确定一个初始划分，然后对初始划分进行优化。这个初始聚类中心的选择对聚类结果有较大的影响，一旦初始值选择的不好，可能无法得到有效的聚类结果；

d）该算法需要不断地进行样本分类调整，不断地计算调整后的新的聚类中心，因此当数据量非常大时，算法的时间开销是非常大的；

e）若簇中含有异常点，将导致均值偏离严重（即:对噪声和孤立点数据敏感）；

f）不适用于发现非凸形状的簇或者大小差别很大的簇。

NaiveBayes分类缺点：

无法处理基于特征组合所产生的变化结果。

## 5.2模型改进

**1）使用多元线性回归查看所选变量与最终结论之间的相关性。**

################多元线性回归############################

airdatanew2<-airdatanew

airdatanew2$cluster<-sapply(fit\_kmeans\_scale$cluster, as.numeric)

airreg<-airdatanew2[,4:9]

str(airreg)

library(car)

#->num

airreg$cluster<-sapply(airreg$cluster, as.numeric)

#相关系数

cor(airreg)

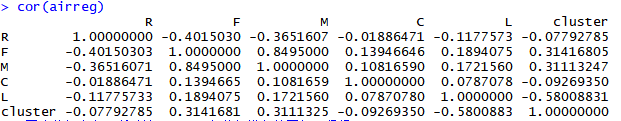


图33.相关系数

从相关系数可以看出，客户最近一次乘坐公司飞机距观测窗口结束的月数与其他变量都成负相关客户在观测窗口内乘坐公司飞机的次数和客户在观测窗口内累计的飞行里程相关系数为0.8495000，相关性较强；客户类别与和会员入会时间距观测窗口结束时间的月数负相关性较强。

# 6结论

**6.1 KMeans聚类模型结论**

表3：KMeans聚类模型结论

|  |  |  |
| --- | --- | --- |
| 客户群 | 排名 | 排名含义 |
| 客户群1 | 1 | 重要保持客户 |
| 客户群2 | 2 | 重要发展客户 |
| 客户群3 | 3 | 重要挽留客户 |
| 客户群4 | 4 | 一般客户 |
| 客户群5 | 5 | 低价值客户 |

#L="LOAD\_TIME"-"FFP\_DATE"(观测窗口的结束时间-入会时间)（单位：月）

#R="LAST\_TO\_END"（最后一次乘机时间至观测窗口末端时间长）（单位：月）

#F="FLIGHT\_COUNT"（观测窗口内的飞行次数）（单位：次）

#M="SEG\_KM\_SUM"（观测窗口的总飞行公里数）（单位：公里）

#C="avg\_discount"（平均折扣率）

L和R越大，表示时间越长。L短表示新客户，意味着要激励，长意味着要挽留，在优劣势分析时，将L越长的设置为优势。C越大，表示折扣越低，那么暗示客户对价格不敏感。M和F越大，表示客户乘机需求越大。

重要保持客户(客户群1)：

这类客户的平均折扣率(C)较高(一般所乘航班的舱位的等级较高)，最近乘坐过本公司航班(R)低，乘坐的次数(F)或里程(M)较高。他们是公司的高价值客户，是最理想的客户类型，对航空公司的贡献率大，但是所占的比例较小。

航空公司应当优先将资源投放在他们身上，对他们做到一对一精准营销，比如提供相应的优惠政策，提高这类客户的忠诚度和满意度，尽可能延长这类客户的高消费水平。

重要发展客户(客户群5)：

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，乘坐的次数(F)或里程(M)较低。这类客户入会时长(L)短，他们是公司的潜在价值客户。

航空公司在运营过程中也要相应加大对这部分客户的投入，比如以短信和邮件的方式询问客户是否满意、提供适当的优惠服务等，加强这类客户的满意度，使他们逐渐成为公司的忠实客户。

重要挽留客户(客户群2)：

这类客户的平均折扣率(C)、乘坐的次数(F)或里程(M)较高，但是已经较长时间已经没有乘坐过本公司航班，即(R)较高或者乘坐的频率变小，他们是公司的重要挽留客户。

航空公司在运营过程中要积极推测这类客户的异常情况，进行竞争分析，观察其他航空公司有什么营销策略，然后采取有针对性的营销手段，延长客户的生命周期。

一般和低价值客户(客户群3、4)：

这类客户的平均折扣率(C)较低，最近乘坐过本公司航班(R)低，乘坐的次数(F)或里程(M)较低。入会时长(L)也较短，因此可能只是在机票打折的时候才会乘坐本航班。

重要发展客户、重要保持客户、重要挽留客户可以称作为客户生命周期管理中的发展期、稳定期、衰退期三个阶段。

模型应用：

1. 会员的升级和保级。对那些接近但尚未达到的高消费客户采取一定的促销活动，刺激他们通过消费达到相应的升级标准。
2. 首次兑现。对那些接近但尚未达到首次兑现机票的会员，对他们进行提醒，使他们达到首次兑现标准。
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。

**6.2 NaiveBayes分类模型结论**

本文第一个Kmeans聚类算法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。为解决以上问题采用了NaiveBayes分类算法对客户进行分类，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。采用分层抽样随机选取数据的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）。

本文将数据挖掘中的聚类分析方法和决策树分析方法引入航空公司的客户流失分析，主要是提供航空公司客户分类以及流失分析的思路和方法；其次，选取了具体数据训练集进行了分析实验，实验证明该分析方法是有效的，可以根据聚类对客户进行分类并用决策树分析结果得出航空公司客户流失的一些规则，基于流失客户的一些共同特征，航空公司不仅可以提出相应对策和改进措施来保留客户、预防潜在客户流失，而且还能通过提高航空公司竞争力来博得流失客户的信赖，提高他们的满意度和忠诚度，使他们重新成为航空公司的消费者。

当然，在航空公司客户流失分析实际操作中，选取一个小的数据训练集得出的分析结论是缺乏说服力的，此时我们面临的将是成千上万条的数据，数据属性也会达到十几甚至几十个。因此，我们再也不可能用手工来进行计算，取而代之的是计算机软件，SAS、AnswerTree是两种应用较为广泛的决策树分析软件。其次，航空公司在客户流失管理中应注重以下几点:

1.决策者应高度重视

航空公司决策者要把避免客户流失作为客户服务工作的重点，制订不同时期的客户维护战略。积极转变经营策略，以宏观的客户维护战略指导客户服务部门做出具体的维护对策，并定期进行有关客户的重复购买率、需求满足感、对本公司及竞争对手品牌的关注程度等指标的统计工作，以便监督、检测客户维护的实施情况。客户流失管理也必须获得到公司各个层面的支持。管理层要赢得这样的支持，就必须对员工进行培训，并将客户流失率作为主要的绩效考核指标。

2.加快建立客户服务中心

区别于客户咨询电话，对航空公司来说，客户服务中心能扩大服务范围，扩大公司影响，减少营业费用，并且对稳固航空公司最重要的客源有着重要的辅助作用；客户服务中心能改善航空公司服务质量、创造和提升航空公司的品牌优势、优化航空公司的服务流程、开辟新的收入来源、提升信息化的水平。国外几乎所有发达国家的航空公司都在用客户服务中心来增强市场竞争力，扩大销售渠道，提高对常旅客的服务能力，建立良好的公众形象。我国的客户服务中心的建立还处于发展阶段，有待于加快建设的步伐。

3.加强客户关系管理

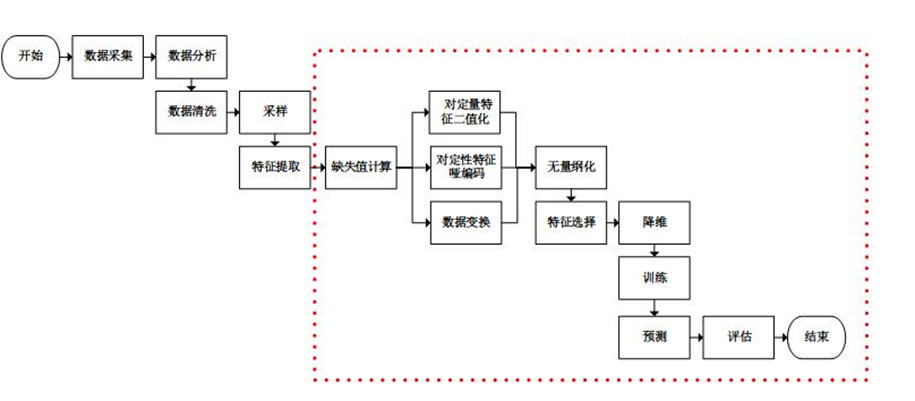
定期收集客户信息，建立客户数据库，获得客户数据是进行客户流失分析的基础。其次，航空公司的客户非常符合托累多的20/80原则，20%的旅客为公司做出了80%的贡献，观察我们身边的人就知道，坐飞机跑来跑去的就那么些人。所以应利用数据挖掘技术，为航空公司找出那20%的旅客，并提供个性化服务是航空公司客户关系管理的关键。其次还应定期对所服务的客户结构进行调查，对最有价值客户进行跟踪并及时反馈，重视最有价值客户的动态管理和预警监管。

总之，利用决策树分析方法对航空公司流失客户进行分析，这对于航空公司加强客户关系管理，保留最有价值客户，提高公司利润增长率都有重大意义。

客户流失对于航空公司造成的损失已日益严重，因此针对航空公司客户流失的分析必须引起航空公司的足够重视。本文将决策树中的NaiveBayes算法引入航空公司的客户流失分析，随机抽取了2/3的样本数据作为训练集来训练建模，结果证明该分析方法是有效的，并且根据测试集对模型的验证可知，模型具有较好的准确性和预测性。最后，根据分析结果，确定了客户流失的共同特征，航空公司可以根据最后一次乘机时间至观察窗口末端时长来设置老客户召回日，从而减少客户流失比例，对于提高航空公司利润增长具有重要意义。

# 7系统收获

1）通过这次的课程设计使我更加深入地了解了数据挖掘与分析的流程与目的



1. 更加了解了这次的课程设计中用到的数据挖掘算法。
2. Means算法的思想很简单，对于给定的样本集，按照样本之间的距离大小，将样本集划分为K个簇。让簇内的点尽量紧密的连在一起，而让簇间的距离尽量的大。 朴素贝叶斯分类是一种十分简单的分类算法，叫它朴素贝叶斯分类是因为这种方法的思想真的很朴素，朴素贝叶斯的思想基础是这样的：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。通俗来说，就好比这么个道理，你在街上看到一个黑人，我问你你猜这哥们哪里来的，你十有八九猜非洲。为什么呢？因为黑人中非洲人的比率最高，当然人家也可能是美洲人或亚洲人，但在没有其它可用信息下，我们会选择条件概率最大的类别，这就是朴素贝叶斯的思想基础。

在回归分析中，如果有两个或两个以上的自变量，就称为多元回归。事实上，一种现象常常是与多个因素相联系的，由多个自变量的最优组合共同来预测或估计因变量，比只用一个自变量进行预测或估计更有效，更符合实际。

3）使我更加熟悉R语言的操作。

了解到R是用于统计分析、绘图的语言和操作环境，是属于GNU系统的一个自由、免费、源代码开放的软件，是一个用于统计计算和统计制图的优秀工具。

# 8参考文献

［1］张宇，张之明．一种基于C5．0决策树的客户流失预测模型研究［J］．统计与信息论坛，2015，30(1):89-94．

［2］王未卿，姚娆，刘澄．商业银行客户流失的影响因素———基于生存分析方法的研究［J］．金融论坛，2014，19(1):73-79．

［3］桂现才，彭宏，王小华．基于决策树的保险客户流失分析［J］．计算机工程与设计，2005，26(8):2026-2029．

［4］张莹．数据仓库在邮政快递企业客户流失分析中的应用［5］李伟健．面向主动服务的客户流失预测模型研究［D］．

［6］邹竞，谢鲲．C4．5算法在移动通信行业客户流失分析中的应用［J］．计算技术与现代化，2009，28(3):98-101．

［7］薛薇．Ｒ语言数据挖掘方法及应用［M］．北京:电子工业出版社，2016:142-143．

［8］吴桐水，贺亮．基于决策树的航空公司客户流失分析［J］．商业现代化，2006(35):381-383．

［9］林楚泉，彭鸿鑫，陈育兴，等．基于ID3算法的航空客户流失模型［J］．大学教育，2013(23):70-72．

# 附录

**全部代码**