2018年MathorCup大学生数学建模挑战赛论文答辩

天津工業大學
TIANJIN POLYTECHNIC UNIVERSITY

基于粗糙集改进的决策树手机精准营销模型

答辩人：张洋 周晓玲

指导老师：教练组

赛题综述

模型的准备

随着我国电子商务和移动支付的快速发展，手机已经成为人们必不可少的工具。除了常规的通讯功能外，手机还可以进行购物、支付、娱乐、学习和交流等。因此，选择一个什么样的手机已经成为广大消费者注重要考虑的问题。

手机的选择是因人而异，除了产品价格、外观、性能等产品因素之外，个人的基本，属性特征(性别、年龄、学历、职业或未来职业)以及个人偏好(如网购热度、网购倾向、娱乐喜好等线上行为路径)也尤为重要。

赛题综述

模型的准备

某品牌手机（命名为Surpass)销售总部希望能够了解消费者对Surpass手机的购买意愿，以便能够进行精准营销。为此，他们进行了相关的调查，得到了附件的数据。

附件“附件1 Surpass目标用户数据”是一份网上关注（包括在网上搜索，浏览，购买过该品牌的三种行为）Surpass手机的目标用户行为数据。数据包括用户基本特征，行为标签，电商行为，视频行为，触媒行为以及Surpass手机基本参数等7个数据表格。所有数据表格的用户编号是一致的，同一个编号指的是同一个用户。

指标一：用户基本信息

手机的选择因人而异，除了产品价格、外观、性能等产品因素之外，个人的基本特征如性别、年龄、职业、学历也尤为重要。

指标二：浏览总时长

浏览时间越长，说明该网站的宣传力度越大。在考虑投资时，在该网页上的投资就可以越高。

指标三：浏览视频总时长

视频浏览时长较大的用户对于手机的屏幕，电池以及内存的要求可能更高一些，而不经常浏览视频的用户对这些指标的关注度较小。

指标四：广告转化率

广告转化率是指通过点击广告进入推广网站的网民形成转化的比例。广告用户的转化量与广告到达量的比值称为广告转化率。

指标五：购买欲望指数

购买欲望指数用来表示用户对于购买该手机的欲望大小，定义为用户在商务平台上浏览该手机的时间与用户在平台上浏览手机的时间之比。

指标六：浏览次数比

浏览次数比表示用户对该型号手机的浏览次数与用户浏览手机的总次数之比，浏览次数比越大，说明用户对该手机的兴趣越大，则用户购买该手机的潜力越大。

指标七：网页影响度

网页影响度指的是用户浏览该网页对其购买该手机的影响的大小。本文中我们是通过已购买该手机的用户对该网页的浏览次数与已购买该手机的用户对所有网站的浏览次数之比来衡量的。

指标八：用户潜力度

用户潜力度即每个用户购买某种手机的可能性，用户潜力度越大，该用户购买某种手机的可能性也越高。

指标九：手机需求指数

手机需求指数横量的是用户购买手机的需求的大小。本文中手机需求指数是通过用户在商务平台上浏览手机的时长与用户浏览总时长之比来衡量的。

指标十：年龄指数

年龄指数是指用户的年龄对用户购买手机概率的影响。我们发现购买用户年龄总是集中在20-30岁之间。我们给每个年龄段赋予不同的分数，即为用户的年龄指数。

指标十一：学历指数

学历指数是衡量用户学历对购买手机的影响。

问题1：

对附件1的数据进行预处理，并进行描述性统计分析。

（1）异常值（包括缺失值，重复值）的处理

对数据进行了预处理，筛选和删除了表格中的重复信息。考虑到附件中所给数据量较大，在保证信息准确的前提下，我们只考虑所有信息均齐全的用户。为了计算的方便，我们将表格中所给的时长均转化为秒，对表格中包含内容信息较多的列，进行了分类提取。

(2) 数据的标准化

用户编号	目标用户行为标签					
6		77856		101843		9571
7		77856		101843		9571
9	网络活跃指数	5776	网络购物指数	2919	在线视频指数	1054
10		5776		2919		1054

$$x^* = \frac{x - \mu}{\sigma}$$

其中，μ代表所有样本数据的均值，σ为所有样本数据的标准差

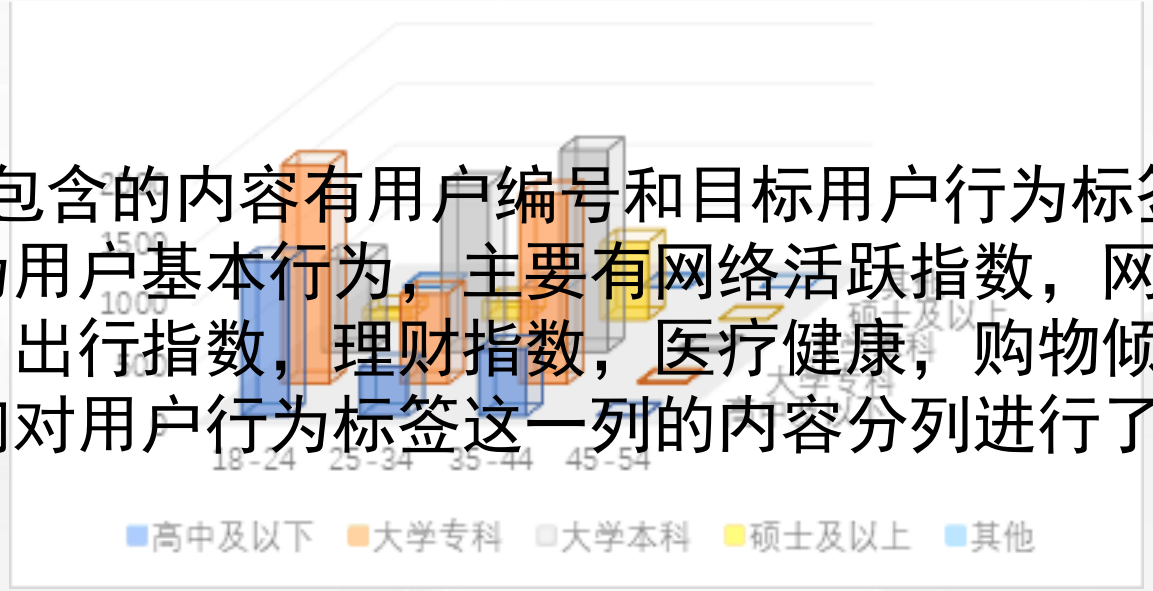
经过这种方法处理的数据符合标准正态分布，得到的标准化后的指标数值

(1) 目标用户

目标用户中所包含的信息有用户编号，在某购物平台上的平均每次停留的时间和最后一次的跟踪状态，其中最后一次的跟踪状态包括购买，搜索和浏览三种情况。我们分别对三种情况下用户的平均浏览时长，浏览次数和浏览比做了统计，浏览比为三种情况下的浏览次数占总次数的比例。具体内容如下表所示:

状 态	平均时长(s)	次 数	频率
购买	1772.911	574	0.032748
浏览	1803.896	11828	0.674806
搜索	1790.783	5126	0.292446
浏览+搜索	1802.311	16954	0.967252

用户编号	网络活跃指数	网络购物指数	在线视频指数
198	1081	268	64
6049	2091	1986	467
560	342	21	181
5718	27208	48821	79



（3）目标用户行为标签

目标用户行为标签中包含的内容有用户编号和目标用户行为标签，其中目标用户行为标签中包含的内容为用户基本行为，主要有网络活跃指数，网络购物指数，在线视频指数，母婴指数，出行指数，理财指数，医疗健康，购物倾向，常用网站，视频网站等。其中，我们对用户行为标签这一列的内容分列进行了提取。

数据的处理

描述性统计分析

用户编号	浏览视频总时长
2	15066s
3	1866s
25	2111s

用户编号	浏览手机总时长	浏览该手机时长	购买手机欲望指数
2	3463s	31s	0.89%
5	5520s	1838s	33.2%
7	1667s	1667s	100%
19	621s	0s	0

（6）触媒行为

触媒行为中包含的内容有用户编号，浏览网页名称，搜索子类名称，网址。将所有购买了该手机的用户对这21类网页的浏览次数进行了统计，访问次数最多，则说明访问该网页对用户购买该手机的影响最显著，以次数从高到低对这21类网页进行排序，分别给其赋从21分到1分，最终求出每个用户的得分和即为网页影响度指标。

数据的处理

描述性统计分析

浏览网页名称	计数	得分
新闻媒体	172194	21
在线视频	129225	20
电子商务	105170	19
搜索服务	105057	18
社交网络和在线社区	76669	17
网址导航	70121	16
网络服务应用	61816	15
IT数码	57337	14
游戏	48457	13
投资金融	34881	12
生活服务	29981	11
汽车	27299	10
音乐	17891	9
房产家居	13703	8
交通旅游	12696	7
休闲娱乐	9141	6
人才招聘	8621	5
医疗保健	6833	4
女性时尚	5762	3
教学及考试	4288	2
垂直行业	1635	1

问题2:

在目标用户中，已经有部分用户在调研期间购买了Supass手机，但更多的用户并没有购买。作为销售部门很想了解用户的基本行为特征对Surpass手机的购买有否影响?如何影响?请你帮助他们解决这个问题。

我们选取了网络活跃指数，网络购物指数，在线视频指数，年龄指数，学历指数，出行指数，理财指数作为用户行为的基本特征。初步考虑这六个指标间的共线性可能比较小，因此我们对选取的6个指标做方差分析，筛选出对是否购买该手机影响显著的前4个指标。再以用户是否该手机为因变量，筛选出的指标为自变量建立了二分类的$logistic$回归模型，得出用户是否购买该手机与用户基本行为特征之间的关系。

Step1:由附件中数据得到x_k, s_k。建立基本方程组

得到因子对应的特征值和因子贡献率。如下表所示：

相关矩阵的特征值：总计=5 平均值=1				
	特征值	差分	比例	累计
1	1.14678638	0.12140784	0.2294	0.2294
2	1.02537854	0.05084284	0.2051	0.4344
3	0.97453570	0.01651845	0.1949	0.6293
4	0.95801725	0.06273511	0.1916	0.8209
5	0.89528214		0.1791	1.0000

通常确定因子个数时，要求因子累计贡献率大于80%，所以我们应该选取4个因子，记为F1,F2，F3,F4。

Step2：确定因子载荷阵系数，得到初始的特征向量。

因子模式				
	Factor1	Factor2	Factor3	Factor 4
网络活跃指数	0.45946	0.36481	0.17397	0.77246
年龄指数	-0.67597	-0.10761	0.04380	0.20971
学历指数	-0.04492	0.79729	0.35633	-0.45063
出行指数	0.33941	-0.49099	0.76073	-0.16597
理财指数	0.60127	-0.06302	-0.48649	-0.29449

得到两个因子对应于指标的模式表，但是由于对应实际问题，公共因子的实际意义不好解释。因此考虑将指标的系数极值化，即让系数趋于0或1，趋于1说明公共因子与该指标密切相关，趋于0时，说明相关程度很低。因此我们做了因子旋转实现系数的极值化。

Step3：方差极大正交旋转，对变量系数极值化（尽量趋于0或1），因子旋转程序运行结果如下：

旋转因子模式				
	Factor1	Factor2	Factor3	Factor4
网络活跃指数	0.04139	0.03570	0.98462	0.16064
年龄指数	0.66472	0.13977	0.15821	0.16733
学历指数	0.00506	0.98320	0.00819	0.03081
出行指数	0.03033	0.03142	0.01058	0.98005
理财指数	0.80376	0.12855	0.03584	0.12803

$$Factor1 = -0.022d1 - 0.059d2 + 0.012d3 - 0.211d4 - 0.751d5$$
$$Factor2 = -0.027d1 - 0.144d2 + 0.979d3 - 0.009d4 - 0.123d5$$
$$Factor3 = 0.983d1 - 0.112d2 - 0.026d3 - 0.026d4 - 0.141d5$$
$$Factor4 = -0.211d1 - 0.112d2 - 0.026d3 - 0.026d4 - 0.751d5$$

对每个因子对应于每一指标的系数进行比较，我们发现第一公因子F1主要体现年龄指数和理财指数，第二公因子F2主要体现学历指数，第三公因子F3主要体现网络活跃指数，第四公因子F4主要体现出行指数。根据以上得到的因子得分函数，可以计算各个样本各个因子的两个样本的得分。所以年龄指数和理财指数是影响用户是否购买该手机的最重要的指标。

Step4：得到因子得分函数，计算样本因子得分。下表展示了一部分得分，全部样本因子得分见附件一。

obs	Objects	Factor1	Factor2	Factor3	Factor4
1	81	0.72289	0.471	1.61382	0.22493
2	157	-0.0028	0.19435	0.70249	1.14702
3	171	1.02077	-0.7402	-0.0171	1.10243
4	195	0.60781	0.42975	0.21127	0.07988
5	203	-1.9816	1.39211	-1.1498	-1.0402
6	243	0.07043	-0.6103	1.19076	1.2329
7	286	-1.1761	-0.7066	-1.1092	0.47983
8	329	-0.2	-1.4538	-1.1307	0.35698
9	379	-1.8837	0.54157	0.4636	0.9959
10	387	0.56744	0.04407	-0.0572	-0.5234

(1)*logistic*回归模型的理论基础

由于*logistic*模型的概率只能为0和1，所以我们对其作如下变换：

$$\text{logit(p)} = \ln\frac{p}{1-p}$$

$$\text{当}0 < \text{p} < 1\text{时，} -\infty < \text{logit(p)} < +\infty$$

所以

$$\ln p1\text{-}p=\beta 0+\beta 1x1+\cdots+\beta 1x1+\varepsilon$$

解得

$$P = \frac{\exp(\beta_0 + \sum_{k=1}^{p}\beta_k x_k)}{1+\exp(\beta_0 + \sum_{k=1}^{p}\beta_k x_k)}$$

P的值表示的是结果为1（即用户购买该手机）的概率；β_i 各变量的回归系数,x_i 为第i个解释变量；

(2) *logistic*回归的参数估计

我们采用极大似然估计法，要直接对此式进行求解比较困难，因此我们采用牛顿-拉普森迭代算法（*Newton-Raphson*迭代算法）：即通过分析它的几何意义得到。

(3) 确定主要影响因子及建立模型

我们建立以任务完成情况的0-1变量为因变量，0表示用户未购买该手机，1表示用户购买该手机；以我们之前的定义的因素为自变量，分别为网络活跃指数，年龄指数，学历指数，出行指数，理财指数的*logistic*模型。

(4) 模型的检验：

检验全局0假设：beta=0			
检验	卡方	自由度	Pr>卡方
似然比	36.7085	5	<.0001
评分	27.3509	5	<.0001
Wald	2.3179	5	0.0836

最大似然估计分析					
参数	自由度	估计	标准误差	Wald卡方	Pr>卡方
Intercept	1	24.891	18.2766	1.8548	0.0532
x1	1	0.0019	0.00150	1.5997	0.0059
x2	1	0.0031	0.00244	1.6604	0.0275
x3	1	0.0039	0.00288	1.8342	0.0456
x4	1	0.025	0.0180	2.0197	0.0553
x5	1	0.0030	0.00605	0.2583	0.0013

pr>卡方这一列值中所有的数字都小于0.05，因此说明模型整体拟合较好。

所有*pr*>卡方这一列中是所有的数都小于0.05，因此每一个指标的拟合度都比较好，说明模型合理。

（5）结果的分析

最终得到用户是否购买该手机与网络活跃指数，年龄指数，学历指数，出行指数，理财指数的线形关系为：

$$\ln\frac{p}{1-p} = 0.19x_1 + 0.031x_2 + 0.39x_3 + 0.025x_4 + 0.30x_5 - 24.891$$

我们计算其优势比（比数比）,优势比的计算公式为

$$\mathrm{OR} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

所以得到$OR_1 = e^{\beta_1}$，由此类推，可得其他四个因素的优势比为$OR_2 = e^{\beta_2}, OR_3 = e^{\beta_3}$，$OR_4 = e^{\beta_4}$，$OR_5 = e^{\beta_5}$。$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$为$x_1, x_2, x_3, x_4, x_5$的系数。因为$x_1$到$x_5$的系数都大于0，所有指标的优势比都大于1，说明四个因素均对用户是否购买该手机有影响。

因此我们得出结论：用户基本行为特征对用户是否购买该手机有影响。

问题3:

一般来讲，不同的网络关注会体现出不同的手机消费个人偏好，导致每个人购买手机的主要动机并不相同。而不同的手机也有着不同的性能。销售部门也很想了解消费者的个人偏好对Surpass手机的购买有否影响?如何影响?同样请你帮助他们解决这一问题。

考虑到选取的指标中可能有较强的相关性，因此我们利用主成分分析对指标进行筛选，选出对用户是否购买该手机影响最显著的前三个指标。以用户是否购买该手机为因变量，以选出的三个指标作为自变量，建立二分类的$logistic$回归模型，得到用户是否购买手机与用户手机消费个人偏好之间的关系。最后，对模型参数进行了灵敏度分析。

Step1： x_1, x_2, x_3, x_4为 四个个人偏好指标为：浏览视频总时长，购买欲望指数，浏览次数比以及触媒分类行为。记为$\mathrm{X} = (x_1, x_2, x_3, x_4)^T$，协方差矩阵为 A。

Step2：为找出综合指标，寻求原始变量X_1, X_2, X_3, X_4的线性组合F_i，其数学模型为

$$\begin{cases} F_1 = u_{11}X_1 + u_{21}X_2 + \cdots + u_{p1}X_p = u_1^T X \\ F_2 = u_{12}X_1 + u_{22}X_2 + \cdots + u_{p2}X_p = u_2^T X \\ \cdots \\ F_p = u_{1p}X_1 + u_{2p}X_2 + \cdots + u_{pp}X_p = u_p^T X \end{cases}$$

其中 $\mathrm{F} = (F_1, F_2, \ldots, F_p)^T$, $\mathrm{U} = (u_1, u_2, \ldots, u_p)$。

这个方程组满足以下条件

$$u_{1i}^2 + u_{2i}^2 + \cdots + u_{pi}^2 = 1 \ (i = 1,2, \ldots, p)$$

$$\mathrm{Cov}(F_i, F_j) = 0, i \neq j \text{且} \mathrm{i,j} = 1,2, \ldots, \mathrm{p}$$

$$\mathrm{D}(F_1) \geq D(F_2) \geq \cdots \geq D(F_n)$$

Step3：确定主成分的过程：寻找正交矩阵U使协方差矩阵A对角化的过程

$$\mathrm{A}=\begin{pmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p}\\ \vdots & \vdots & \cdots & \vdots\\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}\end{pmatrix}\rightarrow U^{T}AU=\begin{pmatrix}\lambda_{1} & \cdots & \\ \vdots & \ddots & \vdots\\ & \cdots & \lambda_{p}\end{pmatrix}$$

结论：$\sum_{i=1}^{p}D(F_i)=\lambda_1+\lambda_2+\cdots+\lambda_p=\sigma_1^2+\sigma_2^2+\cdots+\sigma_p^2=\sum_{i=1}^{p}D(X_i)$

Step4：主成分选取

$\frac{\lambda_k}{\sum_{i=1}^{p}\lambda_i}$为主成分$F_k$的方差贡献率

$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{p}\lambda_i}$为主成分$F_1,F_2,\ldots,F_p$的累计方差贡献率

我们可以得出四个主成分与指标之间的具体关系为：

$$
\begin{aligned}
\text{prin1} &= 0.706x1 + 0.149x2 - 0.688x3 + 0.072x4 \\
\text{prin2} &= -0.020x1 + 0.680x2 + 0.201x3 + 0.704x4 \\
\text{prin3} &= -0.006x1 + 0.706x2 + 0.072x3 - 0.703x4 \\
\text{prin4} &= 0.707x1 - 0.123x2 + 0.693x3 - 0.058x4
\end{aligned}
$$

5.3.3.3*logistic*模型的结果

通过表 我们可以看出用户是否购买该手机与浏览视频总时长，购买欲望指数，浏览次数比，网页影响度之间的线形关系为：

$$\ln\frac{p}{1-p} = -0.0002x_1 - 0.4916x_2 + 0.7331x_3 + 0.0044x_4$$

与第二问一样，我们计算其优势比（比数比）,得到 $OR_1 = e^{\beta_1}$ ，由此类推，可得其他四个因素的优势比为$OR_2 = e^{\beta_2}, OR_3 = e^{\beta_3}$，$OR_4 = e^{\beta_4}$，$OR_5 = e^{\beta_5}$。$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$为$x_1, x_2, x_3, x_4, x_5$的系数。因为$x_1$到$x_5$的系数都大于0，所以指标的优势比都大于1，说明四个因素均对用户是否购买该手机有影响。

因此我们得出结论：用户个人偏好对*surpass*手机购买有影响。

最大似然估计分析					
参数	自由度	估计	标准误差	Wald卡方	Pr>卡方
Intercept	1	2.5064	0.8588	8.5165	0.0035
x_1	1	0.0002	0.0001	0.2274	0.0231
x_2	1	0.4916	0.6310	0.4360	0.0023
x_3	1	0.7331	0.5767	0.2037	0.0012
x_4	1	0.0044	0.0053	0.4004	0.0434

通过上表我们可以看出，所有*pr*>卡方这一列中是所有的数都小于0.05，因此每一个指标的拟合度都比较好，说明模型合理。得到的*logistic*表达式如下：

$$\ln\frac{p}{1-p} = -0.0002x_1 - 0.4916x_2 + 0.7331x_3 + 0.0044x_4$$

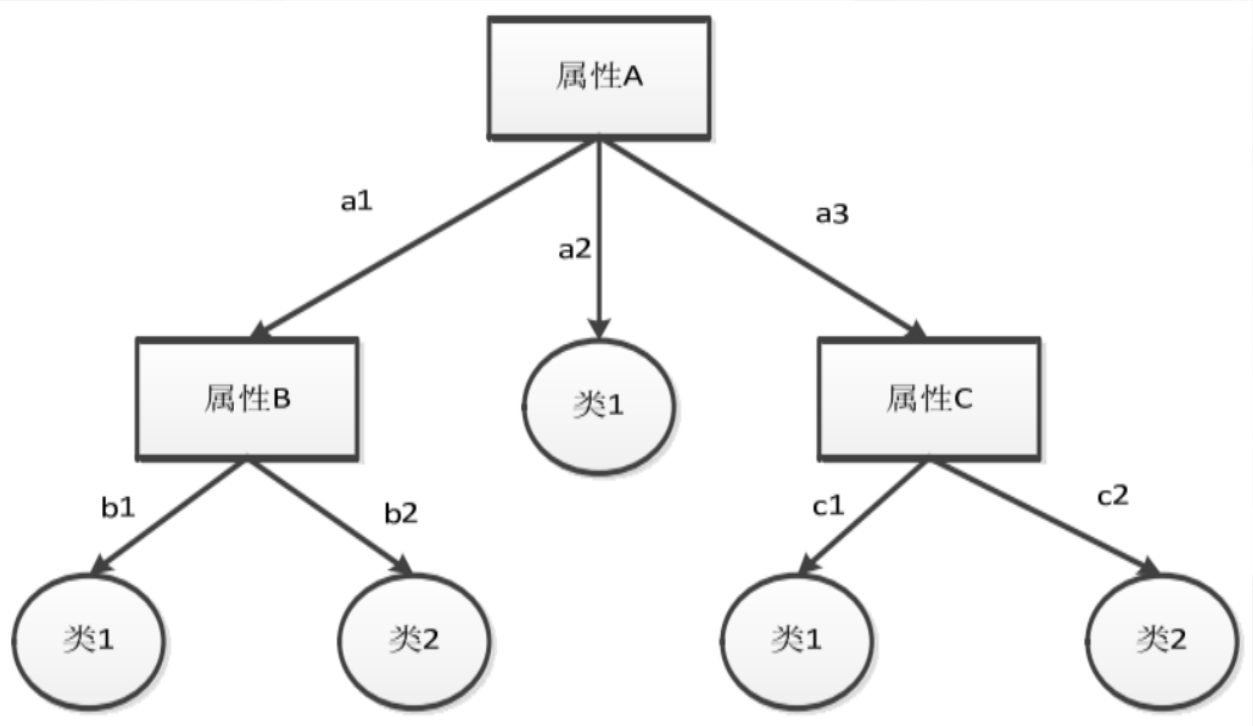
p代表用户购买该手机的概率，x_1代表浏览视频总时长，x_2代表购买欲望指数，x_3代表浏览次数比，x_4代表网页影响度。

问题4:

目前，很多目标用户并没有下单购买Surpass手机，但他们中很有可能有一些是潜在的购买用户。请结合前面的研究,建立一个潜在客户挖掘模型,
对附件2的50位目标用户进行客户潜力界定(需要在正文中展示结果)。运用
你建立的挖掘模型，针对附件1中所有未购买该手机的目标用户，挖掘出100名最有潜力购买Supass手机的目标用户(需要在正文中展示目标用户编号)，并提供建议阐述如何进行精准营销，包括广告投放。

所谓决策树，就是一种树形结构的流程图形，其中根据层次的不同，主要把结点分为根节点、内部节点和叶节点三类。在决策分类中的对一个测试结果采用一个分枝表示，对一个属性的测试就采用节点来表示，而决策树最顶上的那个节点是根节点，叶子节点就代表某一个类或相应类的分布。决策树的生成首先是从根节点开始，然后根据从上到下的递归原理的生成树的一种过程，通过采用分而治之的划分方法将样本数据细分为不同样本来构建模型。

简单的决策树模型如图所示



构造决策树模型之后，下一步就是提取潜在客户的行为特征。在决策树中每个叶节点代表一条规则，即这个规则的左边条件表示根节点开始到达叶节点路径上的全部中间节点组成的一个“与”判断，规则的右边表示叶节点的类型。因此，要对新样本进行分类时，只要该样本数据满足某条分类规则的时，则就可以容易判定它的类（等于规则的右边值）。如果产生的分类规则过多，还需要进一步对这些规则进行处理，合并成更为简洁的形式。

在之前的问题中，我们分析了每个用户的基本行为和个人偏好标签，需要在此基础上挖掘出更多有可能购买该手机的用户进行精准营销。考虑到观看在线视频、浏览网站等行为在很多品牌手机上都可以实现，且该手机屏幕为5.5英寸，我们参考了网上大多数手机的参数，发现*Surpass*手机在并没有突出优势，所以我们放弃了一些与用户体验关联性不高的因素。

综合用户各方面的信息，我们决定采用网络活跃指数、网络购物指数、购买欲望指数、浏览次数之比和性别来判定用户是否会购买Surpass手机。

- **网络活跃指数**

  反映用户上网的频繁程度

- **网络购物指数**

  反映了用户在网上购物的可能性

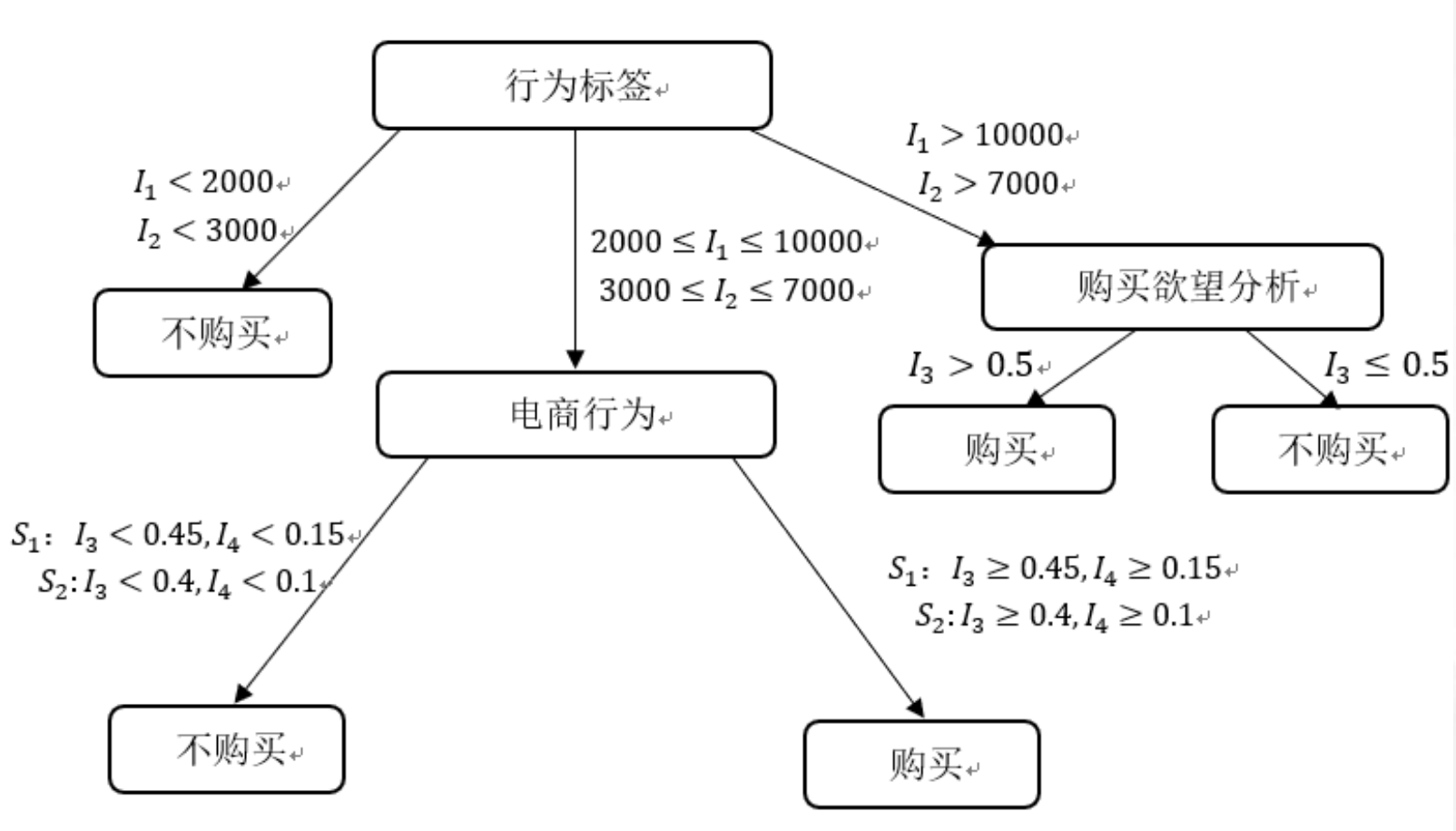
- **购买欲望指数&浏览次数之比**

  表示了用户购买此手机的欲望

- **性别**

  我们分析了购买该手机的用户男女比例，发现女性用户数量要高于男性用户数量，所以我们把女性用户列为优先选择的潜在用户

我们建立并引入了决策树分类模型，如下图所示：



- I_1:网络活跃指数
- I_2:网络购物指数
- I_3:购买欲望指数
- I_4:浏览次数比，
- s_1:男性
- s_2:女性

决策树运行流程如下：

- Step1：判断I_1和I_2，
  I_1<2000, I_2<3000 => 不购买；
  $I_1 > 10000$， $I_2 > 7000$ =>购买欲望分析（Step2）
  $2000 \le I_1 \le 10000$，$3000 \le I_2 \le 7000$ =>电商行为分析（Step3）

- Step2： $I_3 > 0.5$ =>购买
  $I_3 \le 0.5$ =>不购买

- Step3：男性：$I_3 < 0.45, I_4 < 0.15$ =>不购买
  $I_3 \ge 0.45, I_4 \ge 0.15$ =>购买
  女性：$I_3 < 0.4, I_4 < 0.1$ =>不购买
  $I_3 \ge 0.4, I_4 \ge 0.1$ =>购买

对于上述模型中没有考虑到的用户，若用户网络活跃指数较高，网络购物指数较低，则说明该用户平常没有网购的习惯；若用户网络活跃指数较低，网络购物指数较高的用户，说明其平常上网次数不多；若用户购买欲望指数较高，浏览次数比较低，说明该用户仅偶尔几次长时间浏览了该手机，我们猜想此类用户在了解了Surpass参数之后发现该手机不符合自己的需求，所以不再关注该手机；若用户购买欲望指数较低，浏览次数比较高，说明该用户虽经常浏览Surpass手机，但并未长时间浏览。以上用户都不符合潜在用户的条件，所以我们不将该用户列为潜在用户。

为了验证上述树状图的准确性，我们随机抽取了100组已经购买该手机的用户数据进行检验，结果如下

用户编号	预测结果	实际结果	识别是否准确
81	不会	会	否
157	会	会	是
171	会	会	是
394	会	会	是
454	不会	会	否
553	会	会	是
676	会	会	是
695	会	会	是

通过100组用户的预测结果和现实结果的比对，我们发现预测准确率达到89%。说明该分类的预测正确率还是比较高的。而在现实中电子商务挖掘潜在客户时，获取的实时客户行为数据比较充足，进而将使用更多的客户行为数据建立分类模型，从而会使挖掘模型就更加准确，效果更好。

编号	预测结果	编号	预测结果	编号	预测结果
20	购买	2542	购买	5134	不购买
145	不购买	2685	不购买	5139	不购买
471	不购买	2905	不购买	5145	不购买
474	不购买	2925	不购买	5146	购买
528	购买	3293	不购买	5155	不购买
578	不购买	3450	购买	5165	不购买
619	购买	3470	不购买	5174	不购买
697	不购买	3857	购买	5188	不购买
1006	不购买	4216	不购买	5202	购买
1081	购买	4240	不购买	5219	购买
1130	购买	4380	不购买	5380	不购买
1315	不购买	4659	购买	5387	不购买
1440	不购买	4718	不购买	5500	不购买
1539	不购买	5127	购买	5501	购买
2224	不购买	5128	不购买	5513	购买
2319	购买	5129	不购买	5565	不购买
2518	购买	5130	购买		

6.1 模型的优点

（1）本文在指标选取方面思考较为全面，并通过方差分析和主成分分析等方式对指标进行了筛选，最终得出了对价格和任务完成情况影响最显著的指标。因此，模型较为合理。
（2）模型的建立是按照问题的解决思路进行的，我们先分析和发现现有规律，然后对现有的规律进行评价，根据评价标准建立新模型，层次渐进易于理解。
（3）本模型通过对已买该手机的用户进行分析，总结出规律和模型，可以较好的与实际情况相匹配，增强模型准确性。
（4）本模型假设合理，因此模型建立准确，可以较好的符合实际情况，有较强的应用能力，可以与实际紧密联系，结合实际情况解决问题。
（5）模型的可靠性高，可推广性强，在对精准营销问题的求解上有独到的创新之处。对广告投放的规划较为具体，可应用到实际生活中。
（6）本文在分析粗糙集和决策树算法的可行性及有效性的基础上，提出一种基于依赖度改进的属性约简方法，并在此基础上采用新的区分价值的多变量检验改进决策树构造方法建立潜在客户挖掘模型，大大提高了结果的准确性。该混合挖掘方法便于使用，而且高效，同时能处理海量复杂的行为数据，并且在属性个数比较多或者属性之间相关性较大时，决策树的分类效率及准确率较高。
（7）对建立的判断树，我们进行了检验，得到了我们判断树的正确率为89%，正确率较高。

6.2 模型的缺点

（1）本模型的缺点在于附件中给的信息没有全部利用，因部分指标难以总结，搜索数据时间有限，如果有足够的时间，我们会做适当的数据挖掘工作，尽量把所有信息整合成指标体现。

（2）本文提出的粗糙集融合决策树算法只能采用静态的行为数据提取静态的客户行为规则。无法解决时刻变化的行为数据及行为规则提取的问题。目前电子商务网站每天有成千上万的浏览者，客户数据随着时间变化不断的增加和更新，并且这种增加和更新的速度是非常惊人的，故挖掘算法也应该具备动态扩展性，当客户行为数据发生变化时，原先获得的行为知识能够随时进行更新，而不必再利用所有数据为研究对象，再重新进行挖掘。

6.3 模型的推广

本文构建了基于大数据的手机精准营销方案，采用方差分析和主成分分析对指标进行筛选，使用logistic回归模型建立了用户基本行为特征和个人偏好对用户是否购买手机的影响。（加上第四问的）可以用于其他各类产品的精准营销问题，增加产品的销量，提高产品的利润，对于工厂的良性发展具有积极意义。

- 感谢 团队中队友的团结协作；
- 感谢 中国优选法统筹法与经济数学研究会；
- 感谢 MathorCup 数学建模组委会；
- 感谢 评委老师的辛勤评阅；
- 感谢 全体工作人员的辛勤付出！

谢谢！请各位老师批评指正