基于粗糙集改进的决策树手机精准营销模型

摘 要

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。在考虑用户的基本行为特征和个人偏好的基础上，本文对影响手机的销售情况的指标进了统计和分析，建立了基于粗糙集改进的决策树模型，最终实现精准营销。

针对问题一，我们对附件中所给的数据进行了预处理，删除了重复值，缺失值。然后我们对附件中每一个表格的数据都进行了描述性统计分析，将附件中所给的数据整合成我们需要的指标，对这些指标进行归一化，以便于后续建模和计算使用。

针对问题二，结合用户基本行为信息，我们选取了网络活跃指数，网络购物指数，在线视频指数，出行指数，理财指数作为用户行为的基本特征。筛选出已购买该手机用户的这几项指标值，由于指标之间基本无共线性，而购买该手机用户的这些指标可能有趋同性，趋同性越大，则该指标的影响越显著。我们采用**方差分析法**对指标进行了选取。以用户是否购买该手机为因变量，以筛选后的指标为自变量建立了二分类的***logistic*回归模型**，得到用户是否购买该手机与用户基本行为特征之间的函数关系。为探究这些指标的具体影响，我们每次对其中一个指标微小变化，其他指标不变，将变化前后的回归值进行对比，得到每个指标的因子影响率。因子影响率越大，则该指标对用户是否购买该手机的影响越大。最终我们得到因子影响率较大的指标是网络购物指数和出行指数。

针对问题三，结合电商分类，视频行为，触媒行为，我们定义并选取了浏览视频总时长，购买欲望指数，浏览次数比，网页影响度四个指标，筛选出已购买该手机用户的这几项指标值，考虑到指标之间可能存在共线性，我们用**主成分分析**的方法对指标进行筛选。以用户是否购买该手机为因变量，以筛选后的指标为自变量，同样建立二分类的*logistic*回归模型，得到用户是否购买该手机与用户偏好之间的函数关系，用与第二问相同的方法得到每个指标的因子影响率。最终我们得到因子影响率较大的指标是浏览视频总时长和浏览次数比。

针对问题四，对潜在客户行为属性进行约简化处理后，我们构造基于粗糙集的**改进决策树挖掘模型**。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判断，规则的右边表示叶节点的类型。综合用户各方面信息，我们采用网络活跃指数、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。为了验证树状图的准确性，我们随机抽取 100 名已购买该手机的用户数据进行检验，检验正确率达 89%，说明我们的模型判别正确率还比较高。

运用建立的基于粗糙集的改进树挖掘模型，对附件二中给出的 50 名客户进行了潜力界定。利用改进树挖掘模型，我们对附件一中的用户进行潜力度进行分析，选出了前 100 名潜在客户，具体结果见正文。

结合二、三问，我们知道在基本行为方面，影响度较大的指标有网络购物指数和出行指数，在用户偏好方面，影响度较大的指标有浏览视频总时长和浏览次数比。当用户这些指标较大时，我们可向用户推广*surpass*手机，以实现精准营销。在广告投放方面，我们综合网页影响度和各大网页的用户浏览次数两个指标，得到广告转化率这个指标，广告转化率指标越大，我们对该网站投资也越大。

关键词： 精准营销 *logistic*回归 决策树挖掘模型 主成分分析 方差分析

目录

一、问题的重述

1.1 问题的背景与意义

在当今数字信息化时代，传统的营销手段已经不能满足企业的快速发展要求，面临不断涌现的机遇和日益激烈的竞争，新的营销手段应运而生，其中基于大数据进行精准营销的方法得到越来越多的重视。

精准营销就是在精准定位的基础上，依托现代信息技术手段建立个性化的顾客沟通服务体系，实现企业可度量的低成本扩张之路，是有态度的网络营销理念中的核心观点之一。

现阶段手机上网用户整体呈现稳定增长的趋势。如此庞大的手机上网用户群和稳定的增长势头,为手机广告特别是背景下手机广告的发展提供用户数量保证,也预示着未来中国手机广告市场的巨大潜力。随着电子商务和移动支付的快速发展，手机成为人们生活和工作中必不缺少的工具，因此，选择什么样的手机已经成为广大消费者注重考虑的问题。

1.2 文献综述

目前国内外的一些学者在大数据挖掘分析，客户精准营销等方面都进行了许多相关研究。

大数据方面

21 世纪，随着互联网的飞速发展，全球数据量呈现大爆炸的增长，云计算、物联网等新兴产业的诞生，使得人们对于数据的应用需求也进一步增大

2011 年，麦肯锡在其报告《大数据：创新、竞争和生产力的下一个前沿领域》中指出大数据是指平常的数据库工具无法获取、存储、运营和进行分析的众多数据的集合。

2011 年 12 月，我国工信部在物联网十二五规划中，将处理信息的技术作为 4 项技术创新工程提了出来，其中包含的超大数据存储、数据细项挖掘、智能图像及视频分析等内容都属于大数据技术的核心。

精准营销方面

1960 年，美国营销学大师麦卡锡教授提出了 *4P* 营销市场理论，该理论认为产品、价格、渠道和宣传是市场营销的四个基本要素[1]。

1990 年，美国营销大师罗伯特•劳特朋提出了 *4C* 营销理论，该理论重新定义了市场营销的四个基本因素：消费者、成本、便利和沟通。

1999 年，美国的莱斯特 • 伟门提出了精细化营销的概念，他主张改变传统营销渠道及方式，通过建立客户信息资料来进行客户分类，并通过多种更为直接的渠道及方式来进行营销。

2005 年，菲利普•科特勒第一次正式提出了“精准营销”这一概念，他认为企业需要更为精准、高效、能够评估的营销策划，需要更关注效果的营销宣传策划，还需要投入更多资源在挖掘到目标客户。

2006 年，科特勒在《市场营销原理》中，首次提到了以互联网为基础的精准营销理论。同年，齐渊博在《准确营销》一书中将精准营销描述为准确营销，认为精准营销应符合 “标准”和“确定”两个方面，“标准”就是可以有

效地复制推广。并予以进一步优化升级，“确定”就是要求企业必须对市场有非常深入地了解并能够判断未来的市场走势。

1.3 问题的提出

某品牌手机销售总部希望了解消费者对该手机的购买意愿，以便能够进行精准营销。为此，市场营销部门进行了相关调查，得到了附件的数据。为了对此公司的手机进行精准营销，我们需要建立数学模型解决以下几个问题：

（1）对附件中的数据进行预处理，并进行描述性统计分析。

（2）目标用户中，部分用户在调研期间购买了该手机，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是否有影响？并分析具体是怎样影响的。

（3）不同的网络关注会体现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。不同的手机也有不同的性能。销售部门很想了解个人偏好对手机购买是否有影响？并分析是如何影响的。

（4）目前，很多目标用户并没有下单购买该手机，但他们中存在潜在的买家。请结合前面的研究，建立一个潜在客户挖掘模型，对附件2中的50位目标用户进行客户潜力界定，运用此挖掘模型，针对附件1中未购买该手机的目标用户，挖掘出100名最有潜力购买该手机的用户，并提供建议如何进行精准营销和广告投放。

二、问题的假设

假设1：假设一个用户编号即代表一个用户，不存在多个用户使用同一部手机浏览网页的情况。

原 因：精准营销要做的是根据每一位用户的浏览喜好和内容对其进行推荐营销，因此针对的直接对象是用户，即该手机的使用者。并且考虑实际情况，基本是每个人都会使用自己的手机。

假设2：假设用户所浏览的网页，内容和时间都是用户有意为之，不存在误操作的情况。

原 因：只有用户在有意的情况下所浏览的网页和内容才能反映用户的个人偏好和基本特征，而这种情况在实际情况下又比较少，所以在建模时我们对这种情况不予考虑。

假设3：假设附件中所给的用户基本信息都是准确无误的。

原 因：对用户的分析和分类都是建立在用户基本信息基础上进行的，只有在信息无误的情况下，我们对问题所做的讨论和建立的模型才合理。

三、主要符号的说明

数学符号	具体说明
t_{total}	用户浏览网页的总浏览时间
t_{shipin}	用户浏览视频的总时间
α	广告转化率
b_{index}	购买欲望指数
c_{radio}	浏览次数比

f	网页影响度
γ	用户潜力度
β	手机需求指数
b_1	网络活跃指数
b_2	网络购物指数
b_3	在线视频指数
b_4	出行指数
b_5	理财指数
v(l)	剪枝操作前的分类错误样本个数
e(l)	进行剪枝操作的错误样本个数
n(t)	在节点t处的数据样本个数
N(t)	子数T_t的叶子个数

*其他未标明符号在文中说明

四、模型的准备

4.1 重要名词与指标的定义

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。

（1）性别：女性可能更注重外观和感官上的体验，在应用方面注重购物，视频等；男性主要考虑手机的性能和游戏体验。

（2）年龄：年轻人更注重外观和性能的体验，对价格要求不高，而老年人偏爱价格便宜，操作简便的手机。

（3）职业，学历：不同的职业和学历的用户对于手机的特殊性能的要求也有差异。

指标二：浏览总时长

同一用户可能会浏览多个网页，但其在各个网页上的浏览时间是不一样的，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

每个用户的视频浏览时长是有差异的。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。因此，用户的视频浏览总时长应该作为精准营销的一个重要指标。

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。转化是指网民的身份产生转变的标志，如网民从普通浏览者升级为注册用户或购买用户等。转化标志一般指某些特定页面，如注册成功页、购买成功页、下载成功页等，将这些页面的浏览量称为转化量。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。此文中我们以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。该比值越大，说明用户对该手机的兴趣越大，则用户购买该手机的欲望也越高。

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。用户浏览网页的网页影响度之和约大，用户购买该手机的可能性就越大。

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。该比值越大，说明用户对手机的需求越大，从而购买手机的可能性就越高。

五、模型的建立与求解

5.1 数据的预处理和统计分析

5.1.1 问题一的分析

由于本题附件中所给的数据量较大，为了得到更加直观系统的数据，我们首先对数据进行了预处理。处理的内容包括重复值的剔除，缺失值的增添，异常值的筛选和删除等。

为了直观反映数据和在大量的数据中提取出有效数据，我们对附件中的每一个表都进行了统计分析，得出了可直接用于建模和计算的数据。

5.1.2 数据预处理

(1)异常值（包括缺失值，重复值）的处理

由于本题数据有多种异常值和需要处理的项，因此我们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。同时，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。最后，我们将异常值进行了删除和修改。

重复值举例如下表所示：

表 1：重复值举例

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

从上表可以看出，附件中存在大量重复的数据，因此我们需要对这些数据筛选和删除。对于缺失值，经过筛选，在目标用户表格 17578 条数据中，存在 50 个用户的信息是空白的。对于这些信息缺失的用户，我们不对其进行考虑。

（2）数据的归一化

为了处理数据时的方便和消除量纲，我们对得到的数据采用 *Z-score* 标准化的方法对数据进行了归一化操作。公式为：

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

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

5.1.3 数据的统计分析

（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。具体内容如表 1 所示:

表 2 目标用户统计分析表

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

从表中我们可以看出，购买，浏览和搜索跟踪状态下的平均浏览时间相差不大。这可能是因为快餐文化的时代，用户浏览搜索用的时间都比较少。而不同跟踪状态下的浏览次数却有较大差别，这说明浏览次数这一指标可以较好的反应用户的基本行为。可以看出，浏览的次数最多，其次是搜索，最后是购买，这说明购买是经过深思熟虑后做的决定，符合生活中的实际情况。

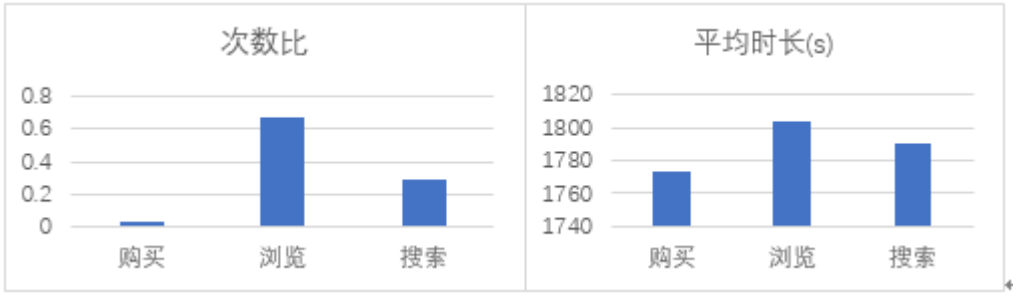


图 1：三种跟踪状态下的次数与平均时长比

由表 2 和图 1，我们可以看出浏览、搜索、购买过手机的用户在购物平台上浏览的平均时长差别很小，证明用户是否购买该手机与用户在购物平台上停留的时长没有很大的相关性，所以我们忽略该因素。但是，我们观察到浏览及搜索该手机的用户频率为 0.967，而真正购买该手机的用户频率为 0.033，也就是说大多数用户浏览过后并没有购买该手机，所以我们认为有必要对目标用户进行精准营销以提升销量。

（2）目标用户身份标签

目标用户身份标签中所包含的内容有用户编号，年龄，性别，学历以及正在学习或从事的专业或职业。对各个年龄段的人数，男生女生的人数，各种学历的人数以及从事不同职业的人数。具体内容见图 2.

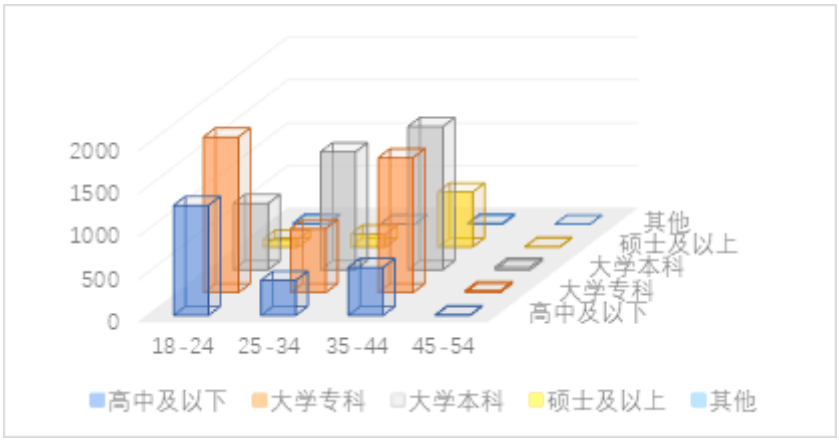


图 2 年龄与学历之间的关系数据图

由图 2 我们可以看出，我们目标用户的结构组成是学历为大学专科和高中及以下的人群中男性居多，而在大学本科及以上的学历中女性较多，且总人数中女性用户要多于男性用户，由此可以得出女性用户倾向于该手机的可能性较大。在 20 到 30 岁左右的人群中，高中及大学本科的人倾向于本款手机的可能性较大，而随着年龄的增长，30-40 岁左右的人群中，大学本科及硕士喜爱本款手机可能性更大。

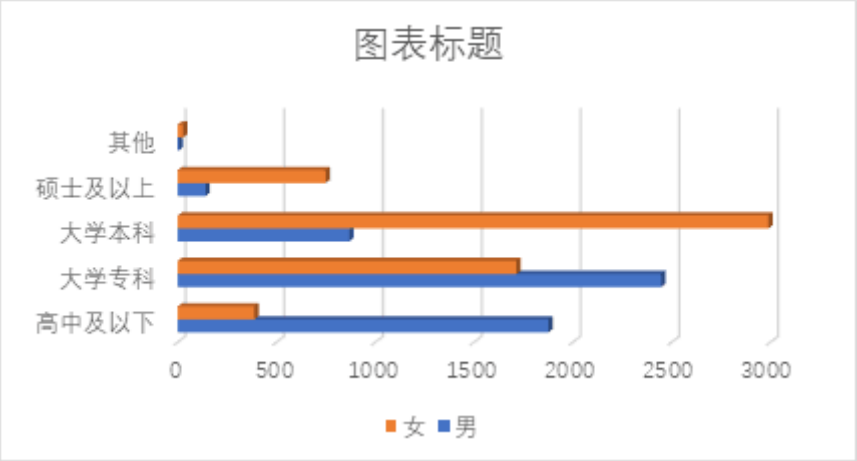


图 3 学历分布情况数据

一般来说学历越高的人年龄就越高，所以我们在网上查阅了中国各阶段学历的人数比例：高中及以下学历人数为 1066406427，占比为 89.9%，大专及以上学历人数为 119636790，占比为%10.1，与上述数据不符，这可能是因为该手机面向的用户都为城市居民，而城市居民的受教育程度普遍偏高，所以该数据与实际现象基本吻合。

（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。提取后的数据如下表所示：

表 3 目标用户行为统计分析表

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

从上表可以看出，购买和未购买该手机用户的网络活跃指数，网络购物指数，在线视频指数的值均有较大差异。因此，用户基本行为特征是影响用户是否购买该手机的重要指标。

（4）电商分类

电商分类中包含的内容主有用户编号，浏览时长，在商务平台网页上浏览过的手机型号，产品一级分类，产品二级分类。我们统计了每个用户浏览手机时长，浏览 surpass 手机时长，计算得到了购买该手机欲望指数，购买该手机的欲望指数为浏览 surpass 手机时长比浏览手机总时长。具体内容见表 4：

表 4：电商行为统计表

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

通过上表，我们可以得到每个用户购买该手机的欲望指数，购买该手机的欲望指数越大，则该用户的购买该手机的潜力越大。

（5）视频行为

视频行为中包含的内容有用户编号，浏览时长，浏览内容，内容一级分类，内容二级分类。考虑到用户观看视频的需求可能对手机的屏幕，内存，电池等有较高要求，并且用户具体观看什么内容与手机本身没有直接关系，我们对每个用户的浏览视频总时长进行了统计，具体内容见表 5：

表 5 视频行为统计分析表

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

通过上表我们可以看出不同用户的浏览视频总时长不同，考虑到 *surpass* 参数中的数据，*surpass* 手机屏幕占比 75.4%，存储 16G，可以满足观看视频较多的用户需要，因此可以作为精准营销时的一个优势。

（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 22 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 22 类网页进行排序，分别给其赋从 22 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，对网页的网页影响度进行统计得到下表所示：

表 6 触媒行为统计分析表

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

通过上表我们看出，不同网页的网页影响度得分是不同的，通过对已购买该手机的用户的基本行为和个人偏好的分析得出的各个网页对用户购买手机的影响力，进而得出的网页影响度指标对于不同用户而言是有显著差异的。网页影响度越大的指标，其购买该手机的可能性就越大。

综上，我们通过数据预处理和对数据的统计分析得到了更为直观和系统的数据，可以用于接下来的建模和计算。对手机的销售情况进行分析，发现影响手机销售情况的因素有很多，例如用户基本行为特征，其中用户基本行为特征又包括网络活跃指数，网络购物指数，在线视频指数，母婴指数，出行指数，理财指数等；用户个人偏好，其包括浏览视频总时长，购买欲望指数，浏览次数比，网页影响度等。要研究手机的销售情况，就要分析手机的销售情况与这些因素变化之间的关系。

5.2 用户基本行为特征对该手机购买的影响

5.2.1 问题二的分析

对于问题二，首先我们从目标用户中筛选出购买该手机的用户，同时我们选取了网络活跃指数，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。

为分析每个指标是如何影响用户是否购买该手机，我们每次给其中一个指标一个微小的变化，将对应的$logistic$回归值作比，得到该指标对用户是否购买该手机的影响的灵敏度大小，从而得到各个指标是如何影响用户是否购买该手机的。

5.2.2 指标的方差分析

（1）方差分析基本思想

方差分析用来研究两个及两个以上控制变量是否对观测变量产生显著影

响。多因素方差分析不仅能够分析多个因素对观测变量的独立影响，更能够分析多个控制因素的交互作用能否对观测变量的分布产生显著影响，进而最终找到利于观测变量的最优组合。

（2）方差分析的主要步骤

Step1:由附件中数据得到x_k, s_k。建立基本方程组：

$$\begin{cases} r_{11}x_1^j + r_{12}x_2^j + \cdots + r_{1m}x_m^j = \gamma_j x_1^j \\ r_{21}x_1^j + r_{22}x_2^j + \cdots + r_{2m}x_m^j = \gamma_j x_2^j \\ \cdots \\ r_{m1}x_1^j + r_{m2}x_2^j + \cdots + r_{mm}x_m^j = \gamma_j x_m^j \end{cases}$$

运行后得到了6个因子对应的特征值，因子贡献率。如下表所示：

表 7 因子贡献率表

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2个因子，记为F1,F2，贡献率分别为。

*Step2：*确定因子载荷阵系数，得到初始的特征向量。

表 8 因子模式表

因子模式				
	Factor1	Factor2	Factor3	Factor4
网络活跃指数	0.45946	0.36481	0.17397	0.77246
网络购物指数	-0.67597	-0.10761	0.04380	0.20971
在线视频指数	-0.04492	0.79729	0.35633	-0.45063
母婴指数	0.33941	-0.49099	0.76073	-0.16597
理财指数	0.60127	-0.06302	-0.48649	-0.29449

通过上表我们可以得到两个因子对应于6个指标的模式表，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：

表 9 正交变换矩阵

	1	2	3	4
1	0.83957	-0.03623	0.43464	0.32387

2	0.02430	0.79909	0.37146	-0.47211
3	-0.39924	0.39135	0.20523	0.80332
4	-0.36760	-0.45497	0.79434	-0.16398

表 10 旋转因子模式表

旋转因子模式				
	Factor1	Factor2	Factor3	Factor4
网络活跃指数	0.04119	-0.00850	0.98452	-0.01034
网络购物指数	-0.66472	-0.13977	-0.15821	-0.16733
在线视频指数	0.00505	0.98320	-0.00819	-0.03081
母婴指数	0.03033	-0.03142	-0.01058	0.98005
理财指数	0.80576	-0.12855	-0.09584	-0.11803

通过上表可以看出，得出的4个因子为

$$\text{Factor1} = -0.022\text{d1} - 0.595\text{d2} + 0.012\text{d3} - 0.011\text{d4} + 0.751\text{d5}$$
$$\text{Factor2} = -0.027\text{d1} - 0.144\text{d2} + 0.979\text{d3} - 0.009\text{d4} - 0.123\text{d5}$$
$$\text{Factor3} = 0.983\text{d1} - 0.112\text{d2} - 0.026\text{d3} - 0.026\text{d4} - 0.141\text{d5}$$
$$\text{Factor4} = -0.027\text{d1} - 0.141\text{d2} - 0.008\text{d3} + 0.977\text{d4} - 0.151\text{d5}$$

对每个因子对应于每个指标的系数进行比较，我们发现第一公因子F1主要体现网络购物指数和常用网站，第二公因子F2主要体现在线视频指数，第三公因子主要体现网络活跃指数，第四公因子主要体现出行指数。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。

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。

表 11 部分样本因子得分表

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

以上为10个样本的4个公共指标的得分，因子F1中171号用户的得分最高，说明该用户的网络购物指数比较大和常用网站指标比较大；因子F2中203号用户的得分最高，说明该用户的在线视频指标比较大；因子F3中的81号用户的得分最高，说明该用户的网络活跃指数比较大；因子F4中的6号用户得分最高，说明该用户的出行指数比较大。

5.2.3 二分类的*logistic*模型的建立与求解

(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，

$$\ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_1 x_1 + \varepsilon$$

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，极大似然估计法是使函数 L(θ)达到最大的参数值θ，作为参数θ的估计值，即取θ使

$$\text{L}(\theta) = \text{L}(x_1, x_2, \ldots, x_n; \theta) = \text{maxL}(x_1, x_2, \ldots, x_n; \theta)$$

$$\text{L}(\theta) = \text{L}(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

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即切线斜率得到。

（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，网络购物指数，在线视频指数，理财指数的*logistic*模型。

（4）模型的检验

以我们最后得到的网络活跃指数，网络购物指数，在线视频指数，出行指数，理财指数指标带入到 *logistic* 模型中得到以下结果：

表 12 模型拟合检验表

模型拟合统计量		
准则	仅截距	截距和协变量
AIC	39.550	12.842
SC	43.910	38.999
-2 Log L	37.550	0.842

从上表可以看出，*AIC，SC,-2LOGL* 对应的截距和协变量值都比较小，因此说明该模型拟合的较好，即指标选取和函数关系都比较合理。

表 13 模型拟合检验表

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

通过上表我们可以看出，*pr>*卡方这一列值中所有的数字都小于 0.05，因此说明模型拟合较好。

表 14 模型参数的最大似然估计表

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

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：

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

p为用户购买该手机的概率，x_1为网络活跃指数，x_2为网络购物指数，x_3为在线视频指数，x_4为出行指数，x_5为理财指数。

（5）结果的分析

通过表 14 我们可以看出用户是否购买该手机与网络活跃指数，网络购物指数，在线视频指数，母婴指数，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

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

为了判断对每个因素的考虑程度，我们对每个因素的数量加 1，如我们先对 x_1 进行加 1，根据 *logistic* 回归模型的公式，我们得到在因素 x_1 增加了 1 以后任务得到完成的概率 p1,得到式（1）

$$\ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \qquad (1)$$

原式为式（2）

$$\ln\frac{p_0}{1-p_0}=\beta_0+\beta_1x_1+\cdots+\beta_px_p \quad (2)$$

用（1）式除以（2）式可得优势比（OR）

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：用户基本行为特征对用户是否购买该手机有影响。

5.2.4 用户基本行为指标如何影响是否购买手机

通过方差分析筛选出用户基本行为指标，通过$logistic$回归分析建立了用户是否购买该手机与这些指标的关系。为了判断这些指标是如何影响用户是否购买该手机的，我们每次改变一个指标，对这个指标增加或者减少 1 个单位，其他指标不变，将改变前后的 y 值做比，得到的比率记为因子影响率。

$$r_i=\frac{y}{y_0}$$

其中，y 为改变某个因子后的 y 值；y_0为未改变该因子时的 y 值；r_i为因子影响率。

因子影响率越大的指标对用户是否购买该手机的影响越大。每个因子的因子影响率如下表所示：

表 15 各个指数的因子影响率

指标	模型原值	变化后的模型值	因子影响率
网络活跃指数	10.57	8.92	0.84
网络购物指数	5.85	10.59	1.81
在线视频指数	9.00	7.56	0.84
出行指数	5.22	9.49	1.82
理财指数	10.16	10.34	1.02

从上表我们可以看出，网络购物指数和出行指数指标的因子影响率较大，网络购物指数的因子影响率大说明用户在网上购买东西的次数越多，则他购买我们手机的可能性就越大。出行指数可能是因为出行指数指的是用户外出购买东西的次数。而网络活跃指数，在线视频指数的因子影响率较小，是因为网络活跃指数大的用户上网比较多，但购买东西并不是很多，而在网上看视频对用户购买手机的影响不大。

所以从精准营销的角度来说，当一个用户的网络购物指数和出行指数较大时，我们便可以向其推销我们的手机，这种情况下潜在客户变成真实客户的可能性较大。

5.3 消费者个人偏好对手机购买的影响

5.3.1 问题三的分析

对于问题三，在问题一数据预处理的基础上，我们选取手机消费个人偏好

指标为：浏览视频总时长，购买欲望指数（购买欲望指数为用户浏览该手机的时长比用户浏览手机总时长），浏览次数比（浏览次数比即为某用户浏览该手机的次数与浏览所有手机的次数之比）以及触媒分类行为。

我们将所有购买了该手机的用户对这 22 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 22 类网页进行排序，并建立了网页影响度指标，网页影响度即为购买了该手机的用户访问该网页的次数比购买了该手机的用户对所有网页访问的总次数。

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。

5.3.2 指标的主成分分析模型

（1）主成分分析的基本思想

主成分分析是对高维数据进行降维的一种方法，将原来的具有共线性的变量重新组合成一组新的相互无关的综合变量来代替原来的变量，从而达到用较少的几个新变量就能综合反映原变量中所包含的主要信息。

（2）主成分分析模型的建立与求解

要分析用户的个人偏好对用户购买手机是否有影响，我们以浏览视频总时长，购买欲望指数，浏览次数比，网页影响度为因素建立主成分分析模型。通过主成分分析模型去除指标间的共线性，具体步骤如下：

Step1： x_1, x_2, x_3, x_4为4个原始指标，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

（3）主成分分析模型的结果

为了确定究竟选择几个特征值，我们得到了主成份分析的相关矩阵的特征值。具体内容见下表：

表 16 主成分分析相关矩阵的特征值

相关矩阵的特征值				
	特征值	差分	比例	累积
1	1.14216908	0.12232552	0.2855	0.2855
2	1.01984355	0.04025669	0.2550	0.5405
3	0.97958686	0.12118635	0.2449	0.8854
4	0.85840051		0.2146	1.0000

从以上表格中，我们可以看出第一，第二，第三主成分的累计解释方差的比率已经超过了 85%，所以我们只需要取前三个主成分。每个主成分的具体表示见下表：

表 17 主成分分析的特征向量

特征向量				
	Prin1	Prin2	Prin3	Prin4
X1	0.706	-0.020	-0.006	0.707
X2	0.149	0.680	0.706	-0.123
X3	-0.688	0.201	0.072	0.693
X4	0.072	0.704	-0.703	-0.058

由上表，我们可以得出四个主成分与指标之间的具体关系为：

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

5.3.3 二分类的 *logistic* 模型的建立与求解

根据问题二模型，仅将改动地方进行如下说明：

5.3.3.1 模型的建立

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；以我们之前筛选后的因素为自变量，分别为浏览视频总时长，购买欲望指数，浏览次数比，网页影响度的 *logistic* 模型。

5.3.3.2*logistic* 模型检验

为检验 *logistic* 模型的准确性，我们对 *logistic* 模型进行了如下检验：

表 18 偏差和 *Pearson* 拟合优度统计量

偏差和 Pearson 拟合优度统计量				
准则	值	自由度	值/自由度	Pr > 卡方
偏差	242.0499	570	0.4246	0.01
Pearson	578.0915	570	1.0142	0.0381

从上表我们可以看出，卡方检验的值均小于 0.05，所以模型通过了斯皮尔曼和方差检验，说明模型整体拟合效果较好。

表 19 模型拟合检验表

模型拟合统计量		
准则	仅截距	截距和协变量
AIC	249.058	252.050
SC	253.412	273.822
-2LOGL	247.058	242.050

从上表可以看出，*AIC，SC,-2LOGL* 对应的截距和协变量值都比较小，因此说明该模型拟合的较好，即指标选取和函数关系都比较合理。

表 20 模型拟合检验表

检验全局 0 假设：beta=0			
检验	卡方	自由度	Pr>卡方
似然比	5. 0078	4	0.042
评分	4.8336	4	0.011
Wald	4.7253	4	0.032

从上表我们可以看出，*pr>*卡方这一列值中所有的数字都小于 0.05，因此说明模型拟合较好。

表 21 模型参数的最大似然估计表

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。

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。

5.3.4 用户偏好如何影响是否购买手机

通过主成份分析筛选出用户基本行为指标，通过$logistic$回归分析建立了用户是否购买该手机与这些指标的关系。为了判断这些指标是如何影响用户是否购买该手机的，我们仍采用第二问的因子影响率的概念来判断因子对用户是否购该手机的影响的大小。

表 22 各个指数的因子影响率

指标	模型原值	变化后的模型值	因子影响率
浏览视频总时长	7.79	7.55	0.97
购买欲望指数	8.58	2.71	0.32
浏览次数比	8.43	8.06	0.96
网页影响度	4.92	1.32	0.27

从上表我们可以看出，浏览视频总时长和浏览次数比指标的因子影响率比较大，这可能是因为浏览视频多的用户，对手机的要求较高，所以会经常购买手机，因此对手机的需求大。浏览次数比属于用户的主观行为，浏览我们手机的次数越多，说明用户对我们手机越亲睐，则越有可能变成实际用户。而购买欲望指数和网页影响度的因子影响率较小。

所以从精准营销方面来说，当一个用户的浏览视频总时长较大和浏览次数比这两个指标较大时，我们就可以向其推荐我们的手机，该用户的购买潜力度也比较大。

5.4 客户潜力界定模型的建立和精准营销

5.4.1 基于粗糙集的改进决策树挖掘模型的建立

根据之前两问的分析，我们找到了影响客户购买的重要因素。接下来在对潜在客户行为属性进行约简化处理之后，我们构造了基于粗糙集的改进决策树挖掘模型。

（1）决策树的生成

目前大多数决策树模型都是采用单变量作为检验属性，就会使得生成的决策树的规模大、分类规则较难理解、决策树存在子树的重复、某些条件属性被多次检验等问题。

本文将采用这种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。首先，给出该算法几个重要的定义：

（2）决策树的剪枝处理

决策树生成完之后，还需要对生成的决策树进行修剪，剪去过细或无必要的分支。目前最常见的决策树修剪重点对象在后剪枝处理上，而重要的后剪枝方法包括有：EBP、MEP、CCP、PEP 等[5]。在这些方法中 PEP（悲观错误剪枝法）是剪枝方法中精度最高的算法之一，而且它还具有剪枝速度快，不需要独立的数据集进行剪枝等优点。综合考虑算法的简单和实用性，本文采用 PEP 剪枝方法对决策树进行剪枝。具体方法如下：

假设存在 $e^t(t) \le e^t(T_t) + S_e(e^t(T_t))$，则子树$T_t$应被剪枝。其中：

$e^t(T_t) = \sum e(t) + \frac{N_t}{2}$；$e^t(t) \le e(t) + \frac{1}{2}$；$S_e\left(e^t(T_t)\right) = \sqrt{e^t(T_t)\frac{n(t)-e^t(T_t)}{n(t)}}$；

$\mathrm{v}(l)$是在剪枝操作前的分类错误样本个数；$\mathrm{e}(l)$表示节点 t 进行剪枝操作而导致的错误样本个数；$\mathrm{n}(t)$表示决策树在节点 t 处的数据样本个数；N_t表示决策树的子树T_t的叶子个数。

（3）提取相应的行为规则

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别（等于规则的右边值）。要是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。

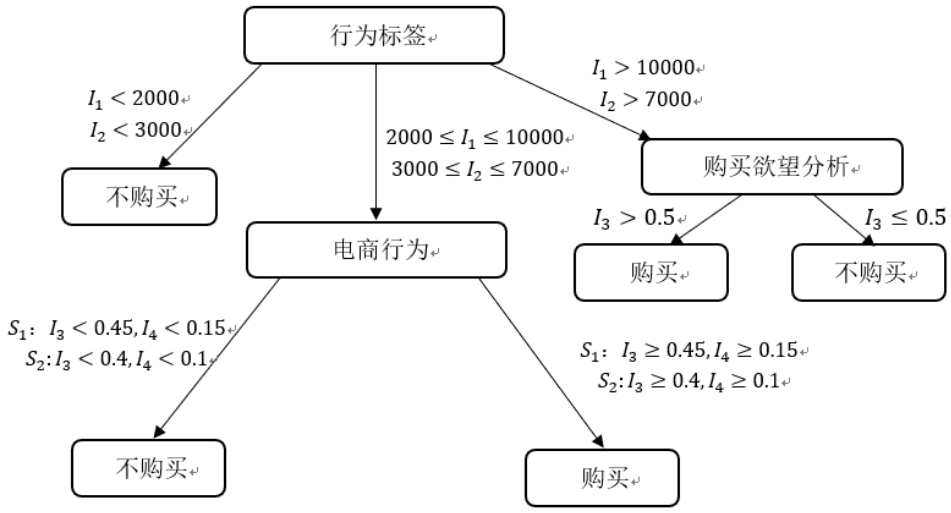
5.4.2 决策树模型分析

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。其中，网络活跃指数反映了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，购买欲望指数与浏览次数之比表示了用户购买此手机的欲望，同时，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

图 4：决策树分类模型

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

其中，I_1代表网络活跃指数，I_2代表网络购物指数，I_3代表购买欲望指数，I_4代表浏览次数比，s_1代表男性，s_2代表女性。

相关步骤解释如下：

Step1：

若用户网络活跃指数小于 2000，网络购物指数小于 3000，说明该用户平时不喜欢上网且不喜欢网上购物，则其不是该手机的潜在用户；若用户网络活跃指数大于 7000，网络购物指数大于 10000，说明该用户很喜欢上网与网购，则直接分析其购买欲望；用户网络活跃指数大于 2000 且小于 10000，网络购物指数大于 3000 且小于 7000，说明该用户较为喜欢网上购物，可进一步分析。

Step2：

在用户很喜欢上网与网购的前提下，若用户购买欲望指数大于 0.5，说明该用户对*Surpass*手机很感兴趣，很有可能购买该手机；若用户购买欲望指数小于 0.5，说明该用户对*Surpass*手机关注度不高，故其不是*Surpass*手机的潜在用户。

Step3：

在用户较为喜欢网上购物的前提下，由于女性用户购买的可能性要略高于男性，所以我们把性别区分考虑：

男性：若用户购买欲望指数大于或等于 0.45，浏览次数比大于或等于 0.15，则为潜在用户；用户购买欲望指数小于 0.45，浏览次数比小于 0.15，则为非潜在用户；

女性：用户购买欲望指数大于或等于 0.4，浏览次数比大于或等于 0.1，则为潜在用户；用户购买欲望指数小于 0.4，浏览次数比小于 0.1，则为非潜在用户。

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。

5.4.3 决策树模型的检验

为了验证上述树状图的可行性，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，其中 20 组预测结果如下表 4-1 所示

表 23 判断树预测结果展示

用户编号’	预测结果	实际结果	识别是否正确
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
752	会	会	是
808	会	会	是
1008	不会	会	否
1010	会	会	是
1062	会	会	是
1071	会	会	是
1112	会	会	是
1136	会	会	是

1163	会	会	是
1180	会	会	是
1225	不会	会	否
1299	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。

5.4.2 50 位目标用户的潜力界定

我们首先定义了用户潜力度指标，用户潜力度即为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。

运用上文中我们建立的基于粗糙集的改进树挖掘模型，我们对附件二中给出的 50 名客户进行了潜力界定，潜力界定的结果如下表所示：

表 24 待判数据分析结果

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		

5.4.3 100 位最有潜力购买用户的挖掘

我们对附件一中没有购买该手机的每一位用户都运用上述建立的基于粗糙集的改进树挖掘模型，从而可以得出每一位用户的潜力指数，将该指数由小到大排序，我们找出了最有潜力购买该手机的前 100 名用户。100 名目标用户编号如下：

表 25 最有潜力购买该手机的用户

100 位最有潜力的用户									
6	10824	47	10972	123	10915	146	171	12787	188
189	195	204	213	458	485	535	329	375	556
11370	10829	621	662	11579	779	891	11741	1030	1032
6056	6562	8898	6753	6543	9987	6745	13121	6819	6892
1054	1064	12964	1112	7518	1149	8378	1241	1246	6939
1259	7732	1389	1395	7442	1653	1682	1880	2139	13316
2388	11301	7665	7635	2904	7354	12650	3012	3069	7060
3125	7772	3186	11064	3207	3217	7329	3322	12042	7171
11231	3471	12407	3615	12205	3823	3832	3997	7331	7156
4056	12450	9993	4280	4288	12139	4398	12608	7276	7233

上表即为附件一中最有可能购买该手机的 100 位用户。

5.4.4 精准营销策略

精准营销就是在精准定位的基础上，依托现代信息技术手段建立个性化的顾客沟通服务体系，实现企业可度量的低成本扩张之路，是有态度的网络营销理念中的核心观点之一。

通过上文对影响用户购买该手机因素的研究，我们可以得到用户是否购买该手机与用户基本行为特征和用户偏好之间的关系，得到影响指标主要由网络活跃指数，网络购物指数，在线视频指数，出行指数，理财指数，浏览视频总时长，购买欲望指数，浏览次数比，网页影响度等并建立了用户上是否购买该手机与这些指标之间的函数关系。函数关系表示如下：

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

p为用户购买该手机的概率，x_1为网络活跃指数，x_2为网络购物指数，x_3为在线视频指数，x_4为出行指数，x_5为理财指数。

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

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。

针对以上两个模型，我们知道在用户基本行为方面，影响度较大的指标有网络购物指数和出行指数，在用户偏好方面，影响度较大的指标有浏览视频总时长和浏览次数比。因此，当用户这些方面的指标较大时，我们就可以向其推广我们的手机，实现精准营销。根据用户这些指标的数据和函数表达式我们可以求得用户是否购买该手机的概率，我们就能及时发现该浏览者的购买意向和购买决策，然后精准地确定其是否是潜在客户，从而可以针对性地展开相应营销策略，争取把更多的潜在用户真正转变为现实用户，增加企业经济效益。实验结果表明本挖掘方法简单、有效和可行，能快速实现电子商务潜在客户的挖掘。

5.4.5 广告投放

我们首先定义了广告转化率指标，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。转化是指网民的身份产生转变的标志，如网民从普通浏览者升级为注册用户或购买用户等。转化标志一般指某些特定页面，如注册成功页、购买成功页、下载成功页等，将这些页面的浏览量称为转化量。

本文中是指由潜在客户转化为实际购买客户的比例。

结合目标用户行为标签和触媒行为中用户对各大网站的浏览量和浏览次数，我们以浏览次数，浏览时间和浏览人数为自变量，以网页影响度为因变量，求出各大网站的网页影响度，根据网页影响度的大小，我们对网站进行了排名，对于排名高的网站，其广告转化率越大，我们对其投资的金额也相应更多。

六 模型的评价与推广

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更高。

（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）本模型的缺点在于运用的方法较为单一，没有运用其他方法对求得的结果进行验证，如果时间充裕的话，可以考虑运用模糊综合评判或主成分综合评价等对问题进行验证求解。

（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。

参考文献

[1]石琦虹，通信运营商借助大数据开展4g手机精准营销的研究 江西财经大学 2017

[2]谢志鹏，张卿，刘宗田.基于粗糙集合理论的决策树生成[J].计算机工程与应用，2000，11：26-28.

[3]苗夺谦，王环.基于粗糙集的多变量决策树的构造方法[J].软件学报，1997，06：26-32.

[4]谢志鹏，张卿，刘宗田.基于粗糙集合理论的决策树生成[J].计算机工程与应2000，11：26-28

[5]王晓平.基于粗糙集的决策树优化算法研究[D].四川师范大学，2013

[6]余春，基于数据挖掘技术的金融数据数据分析系统设计与实现，电子科技大学，2014

[7]任锦鸾，基于大数据的电视节目精准营销，现代传播杂志，2015

[8]杨雨丹,路龙.用户导向的品牌传播应用战略——手机广告的精准营销策略研究[J].品牌研究,2017(06):43-51.

[9]王浩展.基于数据挖掘的高中生手机精准营销策略研究[J].信息与电脑(理论版),2016(19):140-141.

[10]李文辉,王强.手机银行精准营销的适用性分析[J].金融理论与教学,2014(02):45-47.

[11]李克婧,谭浩,王瑞凤.导弹发射瞬时运动安全性分析[J].战术导弹技术,2014(02):28-33.

[12]朱锦宝. 基于智能手机广告的精准营销研究[D].华中科技大学,2012.

[13]王缨.移动VIVA 手机杂志的精准营销[J].中外管理,2011(09):44-45.

[14]桑培铭. 3G背景下我国手机广告精准营销研究[D].华东师范大学,2010.

[15]孟晓佳.浅谈手机广告的精准营销战略和发展前景[J].新闻传播,2009(02):58.