第十二届数学中国数学建模网络挑战赛

地址：数学中国数学建模网络挑战赛组委会　　网址：www.tzmcm.cn

电话：0471-4969085　　邮编：010021　　Email：service@tzmcm.cn

第十二届“认证杯”数学中国

数学建模网络挑战赛

承　诺　书

我们仔细阅读了第十二届“认证杯”数学中国数学建模网络挑战赛的竞赛规则。

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的，如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们接受相应处理结果。

我们允许数学中国网站(www.madio.net)公布论文，以供网友之间学习交流，数学中国网站以非商业目的的论文交流不需要提前取得我们的同意。

我们的参赛队号为：4618

参赛队员　(签名)　：艾茹娟，周春桃，庆淮秀

队员 1：艾茹娟

队员 2：周春桃

队员 3：庆淮秀

参赛队教练员 (签名)：　无

参赛队伍组别（例如本科组）：本科组

地址：数学中国数学建模网络挑战赛组委会 网址：www.tzmcm.cn
电话：0471-4969085 邮编：010021 Email：service@tzmcm.cn

第十二届“认证杯”数学中国

数学建模网络挑战赛

编 号 专 用 页

参赛队伍的参赛队号：（请各个参赛队提前填写好）：

竞赛统一编号（由竞赛组委会送至评委团前编号）：

竞赛评阅编号（由竞赛评委团评阅前进行编号）：

第十二届数学中国数学建模网络挑战赛

地址：数学中国数学建模网络挑战赛组委会　　网址：www.tzmcm.cn

电话：0471-4969085　　邮编：010021　　Email：service@tzmcm.cn

2019 年第十二届“认证杯”数学中国数学建模网络挑战赛第一阶段论文

题　　目　C 题 保险业的数字化变革

关 键 词　车险、Logistic 回归、主成分分析法、层次分析法、UBI 的车险费率厘定模式

摘　　要：

截止到 2018 年底，我国车险规模已达到 3580 亿元，并且车险保费市场复合增长率达到 40%，但这与机动车辆总值在整个社会资产份额中所占比例是不相匹配。但随着信息时代的到来，为车险企业提供了一个更加有力的武器，因此增加续保率对保险企业有着重大的影响。

针对问题一，对附件一中提供的客户进行精准画像，我们以客户、车、车险三类数据为中心，对数据进行降维与提炼，根据相关性原则，选取与目标相关的维度，再对原始数据进行统计分析，精确客户标签画像，最后通过建模知识得到模型标签完善的客户画像。根据作出的精准画像，由于因变量只有续保和不续保两种情况，将以虚拟变量来代替，发生的概率 $P(A=1|x)=p$，不发生的概率即为 $P(A=0|x)=1-P(A=1|x)$，然后建立分组数据 Logistic 回归方程，再根据 *Logit* 变换，建立基于 $x=(x_1,x_2,...,x_n)$ 的 Logistic 线性回归模型，对影响客户续保率的因素定性分析，最后得出续保概率 P。

针对问题二，以从“人”模式与从“车”模式对投保率进行综合分析，采用主成分分析法对影响投保率的多种因素进行降维处理得出主要因素为驾驶行为、车险次数，为改变传统的车险投保模式，我们主要对驾驶行为习惯进行全面的分析，通过数字化技术精确了解客户，制定营销和服务方案，于是基于 UBI 的车险费率厘定模式提出除参考传统车辆因素和驾驶员因素之外，引入驾驶员的驾驶行为习惯。首先，通过对影响驾驶行为的基本因素进行分析，构建对驾驶行为评分指标体系。其次，采用层次分析集成赋权法确定各指标权重，建立基于 UBI 的驾驶行为评分模型。然后，将驾驶行为评分与车险费率之间关联建立挂钩联动模型，给出车险费率调整系数。最后，通过选取具有代表性用户的数据分析，UBI 车险费率对不同安全级别的驾驶行为车主会给予不同程度的保险折扣或增收额外费率。表明基于 UBI 的驾驶行为评分模型比较能准确地反映出驾驶员的驾驶风险水平，所提出的车险费率厘定方法能够实现车险费率的公平化、个性化和差异化，具有较高的实际应用价值。因此基于 UBI 车险费率可以针对不同的客户设计不同的优惠和福利方案，使续保率增加。

参赛队号： 4618

所选题目： C **题**

参　赛　密　码

（由组委会填写）

目录

一、 问题重述

1.1 问题背景

近年来，国际保险行业稳步开展，机动车辆保险在我国的财险保费中所占比重最大，以千亿元计。并且，由于我国汽车保有辆的继续增加和相关车险的政策出台，投保率也呈继续上升趋向。车险一般可占财险公司业务的 70% 到 80%，所以车险市场历来是财险公司的兵家必争之地。以往，财险公司为了赢得市场，往往采取低价、折扣来争抢客户。但是激烈的市场竞争也带来了利润率的下降，甚至有些企业在亏本经营。大多数车企为了提高利润率开始重视承保车辆的质量。重投保车辆质量的做法，其实是险企科学发展的重要体现，是市场竞争下的企业合理行为。中国目前的车险费率制度，大多数符合“从车主义”。即车险保费多少，主要取决于这辆车本身的各项情况，如车的购置价、座位数、排量、购车年限等，根据这些数据计算出一个基本的车险保费价格，再根据这辆车的上年理赔次数来打不同的折扣。这就导致了中国的车险定价模式非常的单调，相似情况的车型，保费也都差不多。已有的定价模式已不能满足新形势的要求，需要制定新型的营销和服务方案。信息时代的到来，为车险企业提供了一个有力的武器，可以通过数字技术来更加精准地了解客户，制定营销和服务方案。

1.2 问题产生

机动车辆保险在我国的财险保费中所占比重最大，并且由于我国汽车保有辆的继续增加和相关车险的政策出台，投保率也呈继续上升趋向。但中国的车险定价模式非常的单调，相似情况的车型，保费也都差不多。为了提高客户续保率，进而提高保险企业的竞争力，需要更加精准地了解客户，制定合理的营销和服务方案。

1.3 研究意义

(一)同过数字技术来更加精准地了解客户，制定营销和服务方案制定可有效增加客户续保率，能够降低车险的经营成本，增加客户在该企业的投保率。

(二)客户资源是保险公司最宝贵的财富源泉,培养忠诚稳定的客户群是公司永续经营的基石,公司如果不注重发展客户,不注重提高客户续保率,无疑会严重影响到公司保费规模的扩大和经营效益的提升。因此，针对客户制定营销和服务方案对于保险公司转变发展方式,实现可持续发展,具有重要的现实意义。

1.4 具体要解决的问题

(一)请建立合理的数学模型，对附件一中提供的客户进行精准画像，给出客户的续保概率。

(二)请针对不同的客户设计不同的优惠和福利方案，以提高续保概率。

二、 问题分析

2.1 问题一的分析

(一)客户精准画像

客户精准画像就是将客户信息标签化，抽象出客户的信息全貌，核心工作就是为客户打标签。本文将对附件一中数据进行预处理，先用标签化客户模型对数据信息进行概括总结，提炼出主要客户特征作为画像标签，再结合提炼出来的标签，应用描述性统计模型对数据进行筛选整合，最后画出分类比例图。

客户续保概率

计算客户续保概率关键在于找到影响续保率的因素，在客户画像的基础上结合与续保率有关的文献先大概确定影响因素，为检验各因素是否影响续保率以及所占权重，本文采用 Logistic 回归模型对续保概率进行求解与检验，最终得到正确的概率计算方程。

2.2 问题二的分析

在此问题中需要针对不同的客户提供不同的车险费率优惠方案以提高续保率。我国目前的车险费率制度，大多数符合“从车主义”，忽视了驾驶行为特征风险。虽每个用户的风险属性不同，但在当前费率模式下计算得到的保费是一样的，这对于驾驶行为良好、驾驶风险事故率低的投保者来说是不优惠的，本次建模把“从人”与”从车“相结合，采用主成分分析与层次分析，得出影响投保率的主要因素。再利用基 UBI 的车险费率厘定遵循“从人+从车"费率模式，制定出对于不同的客户相对于合理优惠的收费标准以提高客户的投保率。

三、 符号说明

符号	说明
$P(A=1\mid x)=p$	根据观测量相对于某客户 x 发生的概率
W	最大特征值对应的特征向量
$\lambda_{\max}$	最大特征值
CI	一致性指标
RI	随机一致性指标
CR	一致性比率
R	表示每个风险单位的纯保费
V	表示变动费用率
Q	表示利润因子
F	表示每个风险单位的固定费用

四、 模型的假设

1. 假设所给数据真实可用。
2. 假设保单性质为转保的客户在保险过期后可能继续采用此种保险。
3. 假设数量较少车种的续保情况可不予考虑。
4. 假设除附表中影响因素外其他因素对续保率影响率较小可忽略不计。

五、 模型的建立与求解

5.1 问题一模型的建立与求解

5.1.1 用户精准画像：描述性统计模型与标签化客户模型

(一)模型的分析

描述性统计模型是一种常用的、简单的数学模型，能够清晰地反映数据特征。标签化模型就是根据对客户的属性、行为等信息的分析精炼出来的特征标识，通过给用户打标签可以概括性的，简明扼要的描述用户特征。对用户标签化处理可以帮助理解用户，方便计算机处理。

(二)模型的建立

以客户、车、车险三类数据为中心，对数据进行降维与提炼。根据相关性原则，选取与目标相关的维度，以避免产生过多数据干扰分析。具体步骤如下：

Step1 数据预处理

由于数据十分庞大，需要先对数据进行预处理，将部分缺失的、错误的、占比非常小的数据，在不影响整体真实性、有效性的前提下进行删除、补充改正或归类。

Step2 基于三个中心逐级给客户的微观画像进行分级

以三个中心为一级分类，再对附件一中所给的数据横向划分得到二级分类，在二级的基础上，对每一种类别数据进行分层次提取精炼得到第三级分类。

Step3 根据步骤一中的分类，对原始数据进行统计分析，精确客户标签画像。

原始输入层主要指客户的历史数据信息，从而达到用户标签体系的事实层，事实层是用户信息的准确描述层。通过建模知识最终得到模型标签：

①原始标签： 用户、车、车险。

②事实标签：客户类型；车辆品牌、价格、车龄、使用性质；购保渠道，出险次数，险种与保费，保额与赔款。

③模型标签：用户性别、年龄；车辆档次、车龄、使用性质；购保渠道与出险次数。

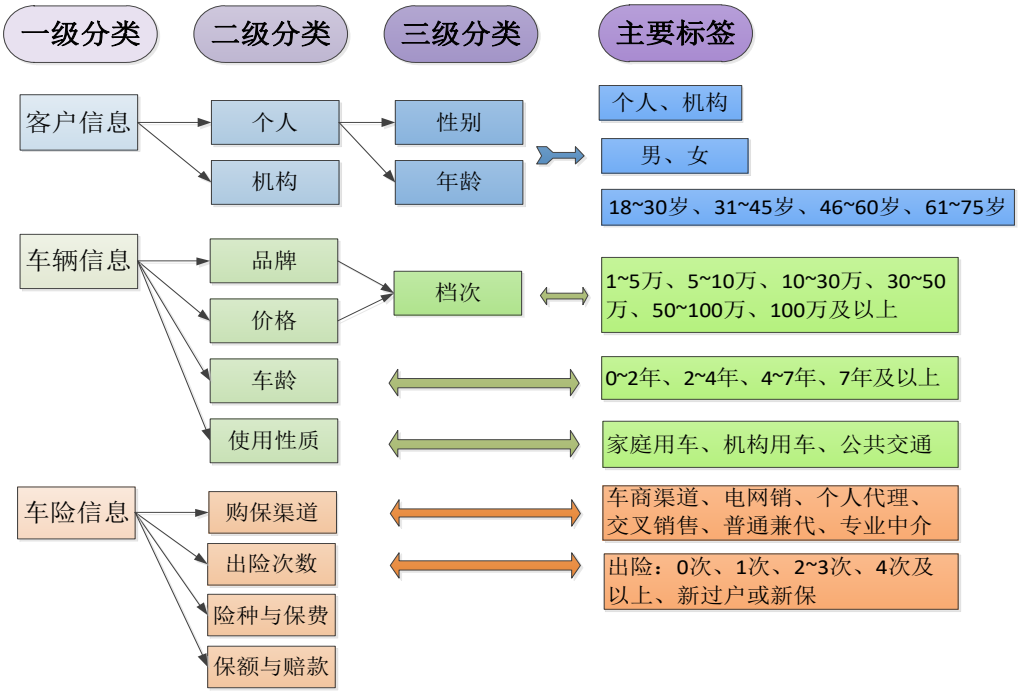
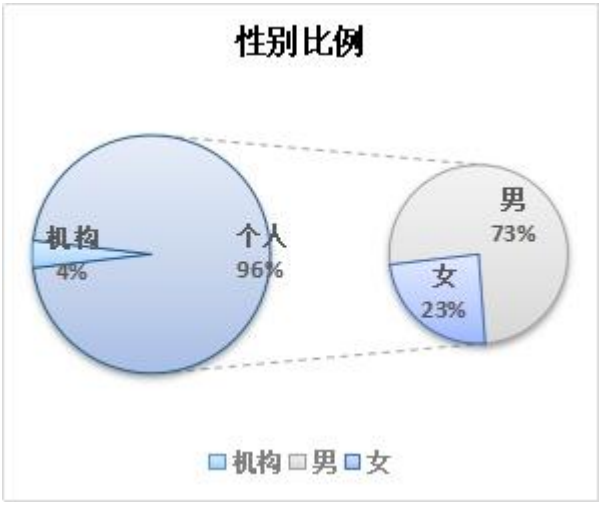


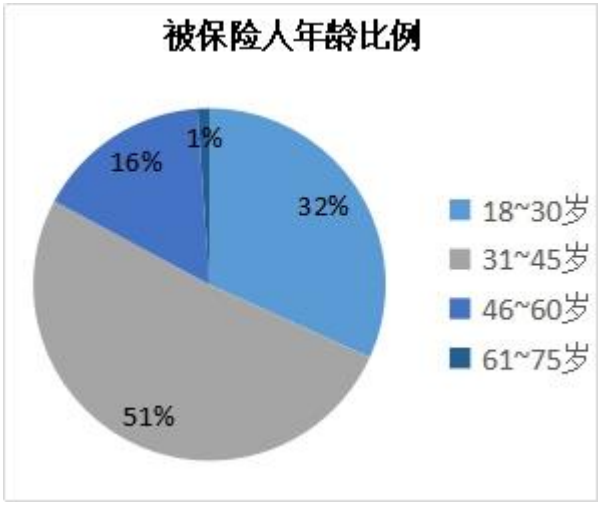
图 1

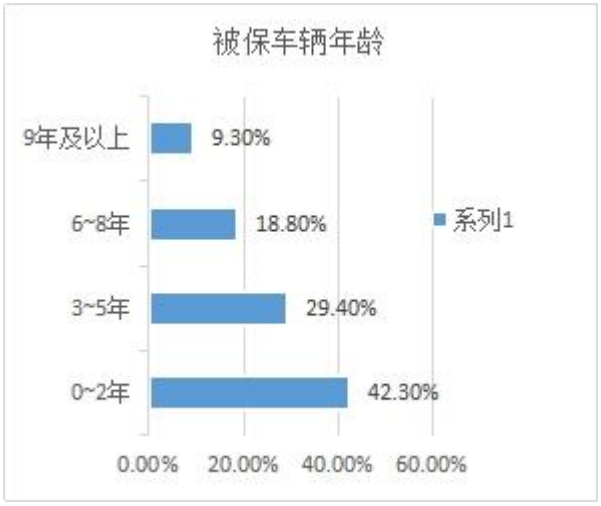
Step4 完善客户画像：根据附件一数据对客户不同标签进行分类统计分析

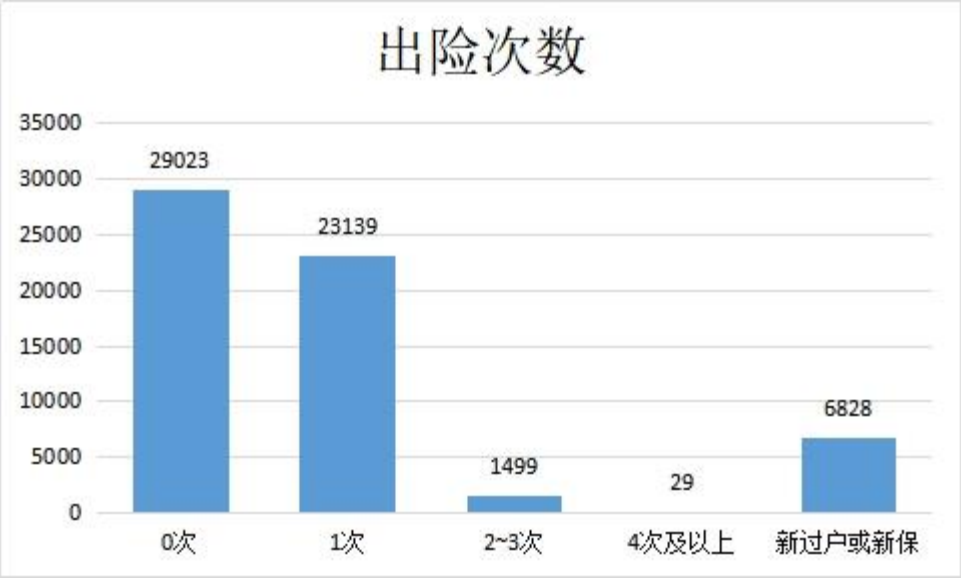
(三)模型的求解

图 2~图 8









5.1.2 客户的续保概率：Logistic 回归模型

(一)模型的选取

线性回归模型是一种常用且有效的数据分析方法，主要将因变量看做由多个自变量所组成的函数，当所有线性回归假设满足时可以用该模型进行预测。但本题中我们所要研究的的续保模型的因变量是典型的二分变量——续保和不续保两种情况，不满足假设，不能简单的用线性回归模型来预测。因此，我们在线性回归的基础上引入 Logistic 回归模型

(二)模型的建立

根据前面对客户精准画像的分析、数据的简单分析以及相关文献资料，从人和车这两个角度考虑，为了避免产生过多不必要的干扰，忽略一些微小影响从多重因素中主要选取了车龄、出险情况、使用性质、新车购价、续保渠道这五个因素作为续保概率的影响因素，以是否续保为因变量，建立分组数据 Logistic 回归方程对影响客户续保的定性分析。具体算法介绍如下：

由于因变量只有续保和不续保两种情况，所以我们可以以虚拟变量来代替，发生续保就是 $A=1$ 的情况，不发生续保就是 $A=0$ 的情况。

假设有 n 个影响续保率的因素，设为 $x_1,x_2,...,x_n$，考虑具有 n 个变量因素的向量 $x=(x_1,x_2,...,x_n)$，设条件概率 $P(A=1|x)=p$（实验指标）为根据观测量相对于某客户 x 发生的概率，则在 x 条件下 y 不发生的概率即为 $P(A=0|x)=1-P(A=1|x)$，所以事件发生于不发生的概率之比为 $\frac{P(A=1|x)}{P(A=0|x)}=\frac{p}{1-p}$，这个比值称为事件的发生比，简记为 $odds$。由于 p 对 x 的变化在 $p=0$ 或 $p=1$ 之间变化不明显，因此对 $odds$ 取对数，令 $y=\ln\left[p/(1-p)\right]$，此式即称为对 p 的 $Logit$ 变换。

设因素 $x_j(j=1,2,...,n)$ 的变化范围为 $\left[x_{j1},x_{j2}\right]$，分别称 x_{j1} 和 x_{j2} 为因素 x_j 的下水平和上水平，并将它们的算术平均值称作因素 x_j 的零水平，用 $x_{j0}=(x_{j1}+x_{j2})/2$ 表示。

表 1 变量水平表

自然变量	x_1	x_2	…	x_n
下水平	x_{11}	x_{21}	…	x_{n1}
上水平	x_{12}	x_{22}	…	x_{n2}
零水平	x_{10}	x_{20}	…	x_{n0}

因此，设影响 p 的变量为 $x_1,x_2,...,x_n$，根据 *Logit* 变换，可建立基于 $x=(x_1,x_2,...,x_n)$ 的 Logistic 线性回归方程：

$$y=\ln\left[p/(1-p)\right]=\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_nx_n \qquad (1)$$

若试验总次数为 $p_i(i=1,2,...,m)$，根据回归正交法与最小乘二原理可求解一次回归方程系数：

$$\beta_0=\frac{1}{m}\sum_{i=1}^{m}y_i=\frac{1}{m}\sum_{i=1}^{m}\ln\left[p_i/(1-p_i)\right]=\ln\sqrt[m]{\prod_{i=1}^{m}\left[p_i/(1-p_i)\right]}=\bar{y} \qquad (2)$$

$$\beta_j=\frac{1}{m}\sum_{i=1}^{m}x_{ji}y_i=\frac{1}{n}\sum_{i=1}^{m}x_{ji}\ln\left[p_i/(1-p_i)\right]=\ln\sqrt[m]{\prod_{i=1}^{m}\left[p_i/(1-p_i)\right]^{x_{ji}}}\quad j=1,2,...,n \qquad (3)$$

(三)实例分析与模型的求解

先对附件中的数据进行预处理，合理的删除，填补，对数据进行分层。再根据上述 Logistic 回归模型对每位客户是否续保的相关数据进行分析与总结，以得到续保率和用户年龄、车龄、出险情况、使用性质、新车购价、续保渠道等因素之间的定性定量关系。

模型中包括车龄、出险情况、使用性质、新车购价、承保渠道这 5 个变量，具体归类如下表：

因素	分层含义
出险次数 x_1	1. 0~1 次；2. 2~4 次；3. 5 次以上
车龄 x_2	1. 0~2 年；2. 2~4 年；3. 4~7 年；4. 7 年以上
使用性质 x_3	1. 私有非盈利；2. 私有盈利；3. 国家公共车；4. 公司营业车
承保渠道 x_4	1. 电网销；2. 交叉销售；3. 车商渠道
新车购价 x_5	1. $[0,5)$万；2. $[5,10)$万；3. $[10,30)$万；4. $[30,50)$万；5. $[50,100)$万；6. $[100,+\infty)$万；

表 2 因素分层表

根据上表对附件一中的数据进行分类筛选，将所得的数据导入 *SPSS* 中，*Analyze*→Re *gression*→*n Binary L.*分步进行 Logistic 回归模型分析，得到如下结果：

表 3 回归系数分析

	回归系数	标准差	df	显著性水平
常数项	-2.8998	0.0123	1	0.021
变量 x_1	0.7738	0.0021	1	0.000
变量 x_2	0.8972	0.022	1	0.084
变量 x_3	0.3526	0.0309	1	0.033
变量 x_4	0.4541	0.0214		0.021
变量 x_5	0.2332	0		0.031

由表 3 可以得到公式（2）中的系数，进而得到回归方程：

$$y=\ln\left[p/(1-p)\right]=-2.8998+0.7738x_1+0.8972x_2+0.3526x_3+0.4541x_4+0.2332x_5$$

得到续保率与出险次数 x_1，车龄 x_2，使用性质 x_3，承保渠道 x_4，新车购价 x_5 这五个因素的 Logistic 回归模型。并可以判断出这 5 个因素对续保率影响的主次顺序为：车龄>出险次数>承保渠道>使用性质>新车购价。

(四)模型的检验

利用方差分析对回归方程与回归系数的显著性水平进行检验分析，结果如下：

表 4 似然比检验

效应	似然比检验		
	卡方	df	显著水平
截距	121.000	1	0.001
车龄	2.776	1	0.016
使用性质	12.500	1	0.005
出险次数	11.000	1	0.0012
新车购价	110.000	4	0.1473
承保渠道	11.300	2	0.1432

由此检验可以看出，这几个因素的效应都是显著的，且模型拟合是充分合理的。综合以上所有分析可以得出续保率与出险次数，车龄，使用性质，承保渠道，新车购价这五个因素之间的关系，从系数上可以判断出这 5 个因素对续保率影响的主次顺序为：车龄>出险次数>承保渠道>使用性质>新车购价，且通过模型回归性检验，模型拟合是充分的。

5.2 问题二模型的建立与求解

5.2.1 问题的分析

研究表明开发新客户的费用是保留老客户费用的 5 倍，所以说成功地保留老客户是企业提高经济效益的重要途径. 正因为如此，国内的车险公司日益注重保单的续保率。从而，我们需要针对不同的客户提供不同的车险费率优惠方案以提高续保率。而我国目前的车险费率制度，大多数符合“从车主义”。忽视了驾驶行为特征风险。虽每个用户的风险属性不同，但在当前费率模式下计算得到的保费是一样的，这对于驾驶行为良好、驾驶风险事故率低的投保者来说是不优惠的。因此，本次建模把“从人”与”从车“相结合，采用主成分分析与层次分析，得出影响投保率的主要因素，再利用基 UBI 的车险费率厘定遵循“从人+从车”费率模式，制定出对于不同的客户相对于合理优惠的收费标准以提高客户的投保率。

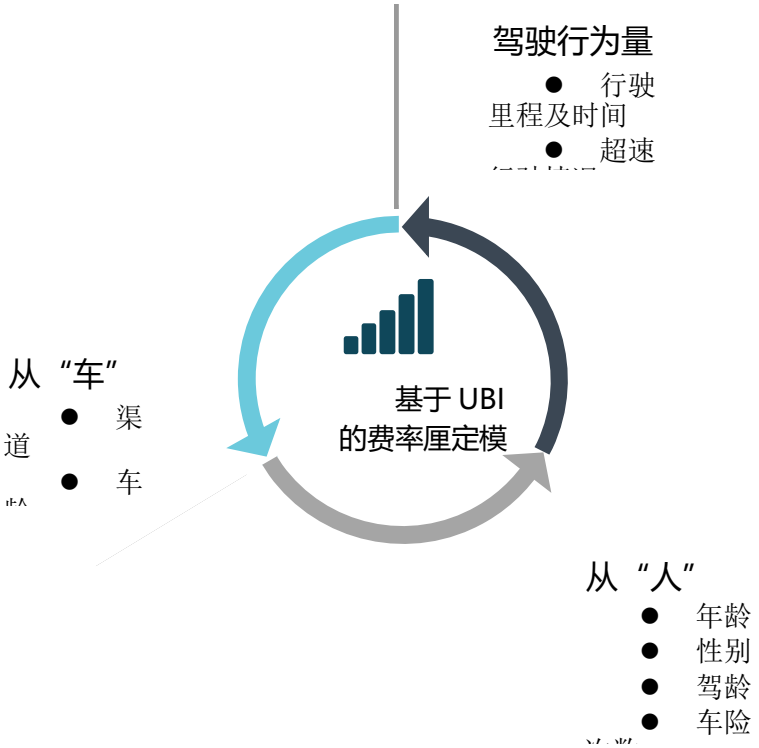


图 2 UBI 的费率厘定模型

5.2.2 基于主成分分析确定影响投保率因素的模型

(一)影响投保率因素的确定

响续保率的因素很多,我们结合附件一以及网上查阅的数据,对其中比较重要的因素进行归纳和总结,将影响续保率的因素从“从车模式”和“从人模式”两个方面划分为 9 个大类，“从车模式”：车龄、新车购置价、使用性质、渠道；“从人模式”：年龄、性别、驾驶行为，驾龄、车险次数。然后根据数据整理，我们把每个大项中的小类进行总结，归纳出几个小类，(根据附件一数据)如下表：

表 5 影响因素分类表

属性名	类数	分类含义
车龄	4	1：车龄 0-2 年；2：车龄 2-4 年；3：车龄 4-7 年；4：车龄 7

		年以上
新车购置价	6	1：[0,5)万；2：[5,10)万；3：[10,30)万；4：[30,50)万；5：[50,100)万；6：[100，+∞)万
使用性质	6	1：党政机关、事业团体用车；2：非营业货车；3：家庭自用车；4：企业非营业用车；5：特种车；6：营业货车
渠道	6	1：车商渠道；2：电网销；3：个人代理；4：交叉销售；5：普通兼带；6：专业中介
年龄	3	1：[18,30)岁；2：[30,45)岁，3：[45,60)岁，4:[60,75)岁
性别	3	1：男(M)；2(F)：女；3：机构(NA)
驾驶行为	3	1:急加速和急减速；2：急转弯；3：疲劳驾驶
驾龄	3	1：[0,10)年；2：[10，20)年；3：[20,50)年
车险次数	3	1：车险 0-1 次；2：车险 2-4 次；3：车险 5 次以上

我们初步分析确定以上 37 个小项作为我们的影响因素，但由于每辆车在车龄，新车购置价，使用性质，渠道以及用户的年龄、性别、驾驶行为、驾龄、车险次数都是确定的，每个大项因素里面的小类都是互斥的，只能满足其中之一。因此与续保率相关影响因素有:车龄(四项)，新车购置价(六项)，使用性质(六项)，渠道(六项)，年龄（三项），性别（三项），驾驶行为（三项），驾龄（三项），车险次数（三项）一共分为九个大类，37 个影响因素。下面对附件一的数据结构进行预处理后，我们根据客观的主成分权数进行分析。

(二)主成分分析的基本概念

主成分分析是把多指标合成为少数几个相互无关的综合指标，即主成分，其中每个主成分都能够反映原始变量的绝大部分信息，且所含信息互不重复。因此通过主成分分析法可以起到降低维度的作用。其中图（）是主成分分析步骤。

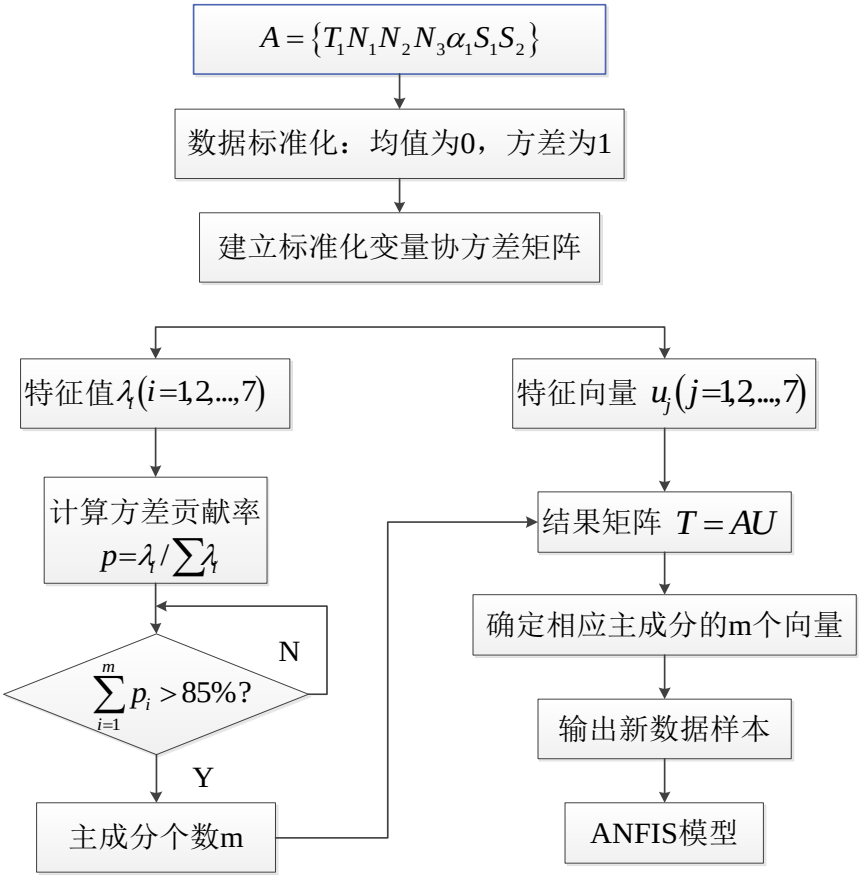


图 3 主成分流程图

(三)建立主成分分析模型

因为投保的车辆数据以及用户人数太多，且投保率受 9 个大因素的影响，因此我们抽取了具有代表性的 6 个用户来研究影响因素之间是否存在相关性，通过数据处理后，利用 SPSS 软件进行目标分析：

(1) 数据标准化

表 6 数据标准化

	渠道	车龄	新车购置价	使用性质	车险	年龄	性别	驾龄	行驶行为
用户 1	-1.668	-1.248	-1.597	-1.555	-1.463	-0.91	-1.57	-0.683	-1.137
用户 2	-0.549	-1.022	-0.771	-0.76	-0.804	-0.566	-0.863	-0.516	-1.015
用户 3	-0.064	-0.142	0.413	0.021	-0.155	-0.892	0.168	-0.42	-0.332
用户 4	0.438	0.31	0.662	0.379	0.483	0.019	0.885	-0.747	0.314
用户 5	0.819	0.863	0.195	0.844	0.779	0.789	0.577	0.577	0.846
用户 6	1.024	1.24	1.099	1.072	1.161	1.56	0.804	1.789	1.324

(2) 系数矩阵的相关性

相关系数矩阵

		渠道	车龄	新车购置价	使用性质	车险	年龄	性别	驾龄	行驶行为
相关系	渠道	1.000	.953	.927	.993	.986	.828	.953	.685	.932

数	车龄	.953	1.000	.892	.978	.986	.892	.917	.785	.994
	新车购置价	.927	.892	1.000	.934	.928	.680	.965	.580	.858
	使用性质	.993	.978	.934	1.000	.994	.834	.957	.713	.957
	车险	.986	.986	.928	.994	1.000	.872	.955	.729	.976
	年龄	.828	.892	.680	.834	.872	1.000	.699	.904	.929
	性别	.953	.917	.965	.957	.955	.699	1.000	.511	.887
	驾龄	.685	.785	.580	.713	.729	.904	.511	1.000	.812
	行驶行为	.932	.994	.858	.957	.976	.929	.887	.812	1.000

a. 该矩阵不是正定矩阵。

表 7 系数矩阵

根据表 7 相关系数矩阵可知，其中 9 个指标之间的相关系数均大于 0.5，可知这 9 个指标之间具有较强的相关性，且车险与使用性质的相关性最高。

(3) 主成分贡献率及累计贡献率

根据 SPASS 作出的主成分贡献率及累计贡献率图可以知道第一个主成分累计贡献率达到 80%以上，即这个主成分就能解释大部分数据，因此可选择这个主成分进行提取分析。

(4) 确定成分得分

根据 SPSS 作出的成分系数矩阵可以得出以下主成分可表示为 9 个指标的线性组合，如下：

$$F_1 = 0.122x_1 + 0.124x_2 + 0.115x_3 + 0.123x_4 + 0.124x_5 + 0.112x_6 + 0.116x_7 + 0.098x_8 + 0.123x_9$$

（4）

由综合得分=主成分得分*方差贡献率，计算可以得到驾驶行为影响因素得分。

再基于 9 个大类中的每个影响因素，我们再选举具有代表性的部分因素，通过数据处理也可以得到小影响因子的相关系数：

表 8 因子相关系数

车险次数	车险 0-1 次	8.2
	车险 2-4 次	12.26
	车险 5 次以上	10.8
车龄	车龄 0-2 年	11.89
	车龄 2-4 年	19.4
	车龄 4-7 年	7.25
	车龄 7 年以上	5.89
新车购置价	[0,5)万	5.02
	[5,10)万	5.21
	[10,30)万	5.56
	[30,50)万	5.46
	[50,100)万	4.68

	[100，+∞)万	4.89
使用性质	党政机关、事业团体用车	12.97
	非营业货车	8.09
	家庭自用车	9.5
	企业非营业用车	10.3
	特种车	13.05
	营业货车	10.1
驾驶行为	疲劳驾驶	15.7
	急加速和急减速	13.2
	急转弯	13.8

从上表中我们可以看出:

1. 驾驶行为习惯越好，风险越低，投保率就低，且还能享受更多的保险优惠政策。
2. 就车龄而言，其中当车龄在 2-4 年时人们的续保的可能性最高。
3. 就车险次数而言，随着使用时间的增加，车的性能下降，出险的可能性就越高。

数据表明，当出险次数在 2-4 次时，人们更愿意续保，而当车为新车或者用时间很长时候人们的对续保的概念性不强，前者主要是车的性能相对较好，人能对车的信任度比较高，而后者则因为车的性能和车的年龄都已经超过了一定的年限，续保的价格相对较高，人们可能会考虑买一辆新车，因此这两种情况人们可能不愿意续保。

4. 就使用性质而言，对安全系数要求较高和使用年限较长的国家公共车和公司营业车相对容易续保。
5. 就新车购买力而言，价位在 20-50 万元的人更容易续保，这些车的车主应该主要在中层阶级，对于购买另外一辆同样价位的车会有很大的负担，另外这些车的使用年限较长，性能相对较好，续保能够对个人而言带来持久效益。而 0-5 万元的车价格相对较低，性能较差，而车险相对较高，性价比不合理，很多人更愿意且有能力买一辆同样价位的车，而不愿续保。

5.2.3 驾驶行为评分指标体系的建立

(一)多层次结构的驾驶行为评分指标体系建立

基于所采集的指标，在遵循基本原则以及借鉴国内外道路交通安全评价影响因素分析的基础上，根据层次分析法的思想建立了多层次结构的驾驶行为评分指标体系，如表所示：

表 9 评分指标体系表

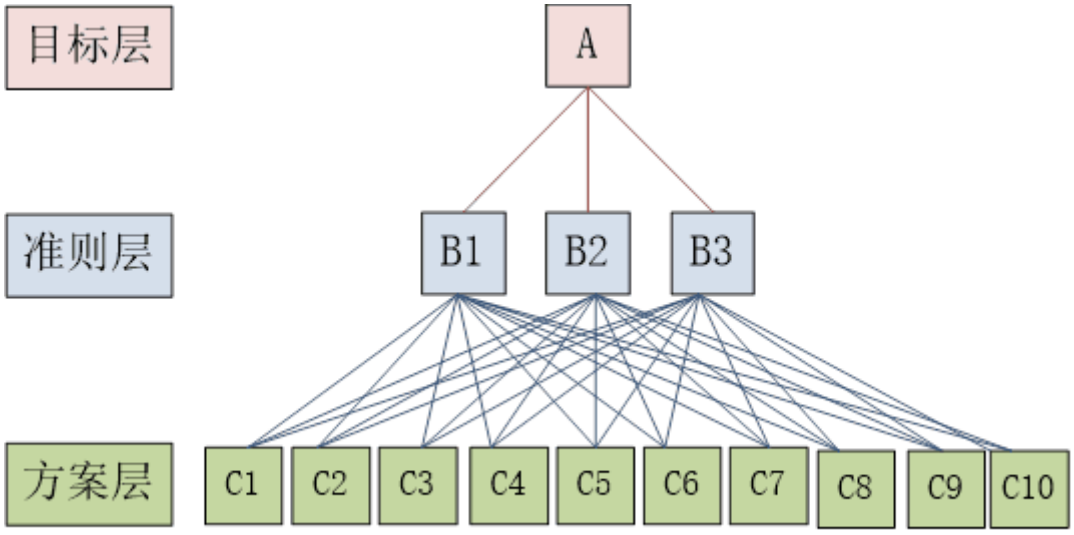
目标层	一级指标	二级指标	说明
驾驶行为评	行驶里程及时间 B1	月总行驶里程 C1	连续变量
		工作日早晚高峰行车时间 C2	连续变量
		夜间行车时间 C3	连续变量
		周末行车时间 C4	连续变量
	超速行驶情况	时速 80～120km/h 行车时间占比 C5	连续变量

分 A	B2 驾车行驶情况 B3	时速高于 120km/h 行车时间占比 C6	连续变量
		急加速次数 C7	连续变量
		急减速次数 C8	连续变量
		急转弯次数 C9	连续变量
		违章次数 C10	连续变量

(二)基于层次分析法对驾驶行为指标权重的确定

(1) 建立层次结构

图 4 层次结构图



(2) 构造判断矩阵并进行一致性检验

表 10 判断矩阵元素标度

标度	含义
1	表示两个因素相比，具有同样重要性
3	表示两个因素相比，一个因素比另一个因素稍微重要
5	表示两个因素相比，一个因素比另一个因素明显重要
7	表示两个因素相比，一个因素比另一个因素强烈重要
9	表示两个因素相比，一个因素比另一个因素极端重要
2，4，6，8	上诉两相邻判断的中值
倒数	因素 与 比较的判断 ，则因素 与 比较的判断

准则层判断矩阵如下

$$A=\begin{pmatrix}1 & 1/4 & 1/3\\ 4 & 1 & 3\\ 3 & 1/3 & 1\end{pmatrix}$$

A 最大特征值：$\lambda_{\max}=3.0735$

相应的特征向量为：$\omega=[\omega_1,\omega_2,\omega_3]=[0.1172,0.6144,0.2684]^T$

随机一致性指标：$RI=1.12$（查表）

一致性比率：$CR=CI/RI=0.0634<0.1$

所以可通过一致性检验

(3) 单层排序一致性检验

类似地，对各层次进行计算：

$$B_1=\begin{pmatrix}1 & 2 & 1/3 & 3\\ 1/2 & 1 & 1/4 & 2\\ 3 & 4 & 1 & 5\\ 1/3 & 1/2 & 1/5 & 1\end{pmatrix}\quad B_2=\begin{pmatrix}1 & 1/5\\ 5 & 1\end{pmatrix}\quad B_3=\begin{pmatrix}1 & 2 & 1/2 & 1\\ 1/2 & 1 & 1/3 & 1\\ 2 & 3 & 1 & 1/\\ 3 & 3 & 2 & 1\end{pmatrix}$$

①行驶里程及时间指标判断矩阵：$w_1=[0.2329,0.1385,0.545,0.0837]^T$

②超速行驶指标判断矩阵：$w_2=[0.1667,0.8333]^T$

③驾车行驶指标判断矩阵计算结果：$w_3=[0.1644,0.1051,0.2848,0.4457]^T$

表 11 各属性的最大特征值：

	行驶里程及时间	超速行驶情况	驾车行驶情况
λ_{max}	4.0511	2	4.0709
CR	0.0189	0	0.0262

由上表可得CR均小于 0.1，因此上述情况均通过一致性检验。

(4) 权重的计算机结果分析

将准则层的权重与指标层的权重进行相乘，得到各指标的最终权重值，结果如下表所示：

表 12 权重计算结果

c	B_1（0.1172）	B_2（0.6144）	B_3（0.2684）	w
c_1	0.2329	—	—	0.0273
c_2	0.1385	—	—	0.0162

c_3	0.545	—	—	0.0639
c_4	0.0837	—	—	0.0098
c_5	—	0.1667	—	0.1024
c_6	—	0.8333	—	0.512
c_7	—	—	0.1644	0.0441
c_8	—	—	0.1051	0.0282
c_9	—	—	0.2848	0.0764
c_{10}	—	—	0.4457	0.1196

由上表可知，采用层次分析法，里程及时间指标占总权重的 11.72%，超速行驶指标占总权重的 61.44%，驾驶行为指标占总权重的 26. 84%。三者的重要程度排序为“超速行驶指标>驾驶行为指标>里程及时间指标”。其中，超速行驶指标在所有指标中占有绝对重要地位，可以看出专家认为超速行驶对驾驶安全影响尤为重要。然而除超速行驶指标之外，其他些重要的指标权重则占的比重较小，如行驶里程指标只占总权重的 2.73%，夜间行车时间指标只占总权重的 6.39%，这些显然与现实是有差距的。虽然超速行驶是引发交通事故的主要原因之一，但里程因素以及夜间行车风险因素也是不能忽略的。由此可见层次分析法具有较大的主观性，所以我们需要对该模型进行改进。

(三) 驾驶行为评分模型的构建

(1) 确定基准安全评分分值及基准评分区间

基准评分分值 P 是衡最驾驶行为是否安全的临界点，基准评分区间是在基准分值的上下土 C 的基础上确定的。在大数据样本的基础上,利用数据挖掘中的特征分析方法建立驾驶行为评分与事故率、赔付率的关联关系，根据大数原则，确定基准评分分值 P 和浮动分值 C。

(2) 确定评分分值区间 Pi(i =1,2,..)

评分分值区问 Pi 的确定是挂钩联动模型的关键。-般而言，在大样本数据的基础上，利用数据挖掘中的聚类分析方法挖掘并分析事故风险与驾驶行为评分的关系，把具有风险相似的驾驶行为评分归为同类，以此确定评分分值区问 Pi。

(3) 车险费率计算

a) 基础费

本次建模采用纯保费法对基础费率进行求解，纯保费是赔款和理赔费用之和。纯保费法(Pure Premium Method)的基本原理是在纯保费的基础上附加各种必要的费用和利润得到保费。纯保费法的费率厘定可通过基本保险方程(Fundamental Insurance Equation)推导得出，其基本保险方程公式如下:

保费=赔款+理赔费用+承保费用+承保利用附加

在费率厘定中，将承保费用进-步划分为固定费用和变动费用两大类。固定费用与保费大小相互独立，对于每个风险单位或每份保单而言是一个常数，变动费用随着保费的变化而变化，与保费成比例。

假定：R 表示每个风险单位的保费；K 表示每个风险单位的纯保费；F 表示每个风险单位的固定费用；V 表示变动费用率；Q 表示利润因子。

则上述保险方程可表示为：

$$\mathrm{R} = \mathrm{K} + \left(\mathrm{F} + \mathrm{RV}\right) + \mathrm{RQ}$$

对上式进行变形，得到如下费率计算公式:

$$\mathrm{R} = \frac{K + F}{1 - V - Q}$$

b) 率费率调整系数

一般地，以基准安全评分区间的费率调整系数为 1，然后根据评分分值区间在费率调整系数浮动范围内给予对应的费率调整系数，在大样本数据的基础上，通过数据挖掘中的关联分析方法、多元统计分析方法建立评分区间与费率调整系数的关联关系，以此来确定费率调整系数 β。

(四)车险费率调整系数的确定

通过查阅资料，我们以 UBI 车险费率对不同安全级别的驾驶行为车主给予不同程度的保险折扣或增收额外费率。UBI 车险费率是以驾驶行为分析和驾驶行为评分机制为基础。综合考虑后，对附件一中 100 名用户的驾驶行为进行评分，假定基准驾驶行为分值为 70 分，驾驶行为分值区间落在基准驾驶行为分值士 5 分，即驾驶行为分值区间为(65,75)，通过计算 UBI 车险费率调整系数与驾驶行为得分间的关系下表所示：

表 13UBI 车险费率调整系数

驾驶行为得分	车费率调整系数
0≤p＜45	1.2
45≤p＜65	1.1
65≤p＜75	1
75≤p＜85	0.9
85≤p≤100	0.8

根据表 11 中车险费率调整系数与驾驶行为得分的关系，分别得出 100 位用户对应的车辆保险费率调整系数值，如图所示：

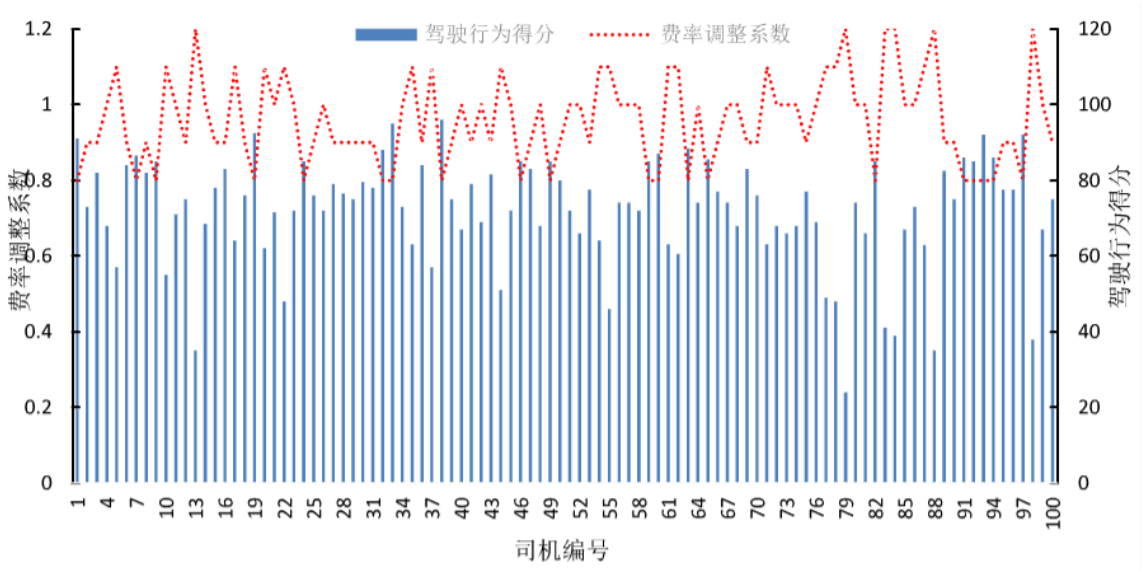


图 5 驾驶行为评分与费率调整系数图

从图 5 中可以看出：

①22 名司机的的保费费率调整系数大于 1,需要支付额外的保费,其中 6 名司机的保费费率调整系数为 1.2，16 名司机的车辆保险费率调整系数为 1.1；

②车险保费费率不需要进行调整的司机有 28 名；

③能获得车险保费优惠的司机有 50 名，其中 30 名司机的保费费率调整系数为 0.9，20 名司机的车辆保险费率调整系数为 0.8。

图 5 中表明：

①50%的司机由于驾驶行为安全评分级别高能够获得 UBI 车险保费费率优惠；

②28%的驾驶驶行为安全评分级别中等的驾驶员的车险费率不进行调整；

③22%的出行者由于驾驶行为表现糟糕需要偿付额外的保费金额

因此，我们以从“车”模式和从“人”模式的影响因素全面分析，针对不同的客户设计不同的优惠和福利方案，使续保率增加。

5.2.4 对不同客户的优惠和福利方案制定

(一)针对出险率低风险小的优质客户

方案一：针对客户有无理赔记录按年数免去不等额的保费，增设绝对免赔额，根据客户需求增设 300 元、500 元、1000 元、2000 元绝对免赔额或者设立保险代金券在客户投保时发放。

方案二：增设多次事故免费率，针对理赔次数不同的客户续保，对于驾龄长、经验丰富、往年理赔较少的车主，可推荐此项进行优惠。

方案三：车损险不足额投保，针对连续几年未出现任何险的优质客户推荐此项目。

方案四：向客户鼓励承保效益险种，通过减少险种达到续保（如减少乘客座位险）。

(二)针对未出险客户

未出险客户一般为优质客户，其在保险期限内没有获得保险理赔，这类客户往往有

一种“交钱后没有得到任何服务”的感觉。当前，公司往往采取下年上保险时的“无赔款优惠”政策留住这类客户，但是优惠不等于服务。这类优质客户群体需要得到公司的特殊关注，提高这一客户群体的满意度，这在提高优质业务续保率中至关重要。制定服务策略如下：

方案一：非保险事故免费道路救援，包括紧急加油、紧急加水、现场抢修、拖车牵引等，以解决客户燃眉之急。

方案二：制定增值服务套餐，如有偿的洗车、打蜡、补胎、四轮定位等。

方案三：代办年检、代办理赔等。

(三)在实际销售过程中，配合前两项，以客户的消费顾虑为基础，抓住客户消费心理，三方面相互结合，及时进行调整。

方案一：险种组合:第三者责任险+车上人员责任险+不计免赔

保障范围是交通事故中造成第三者人和物的损失及车上人员，适用对象为注重基本保障，不想多支出保费的人，且适合各种车型。

方案二：险种组合:车损险+第三者责任险+全车盗抢险+不计免赔特约险+车上人员责任险，保障范围为保车、保人，适用对象为一般公司和个人。

方案三：险种组合:车损险+第三者责任险+全车盗抢险+不计免赔特约险+车上人员责任险+玻璃险+划痕险+修理厂特约

保障范围为最全面的保障车、车上人员，适用对象为机关、事业单位、大公司，特别适合新司机+高档新车。

六、 模型的评价

6.1 模型的优缺点

6.1.1 描述性统计模型与客户标签化模型

优点：

描述性统计模型与客户标签化模型操作简单，容易实现，得到的结果也很直观。

缺点：

虽然该模型操作简单，但是对数据的处理不够精致，得到的结果比较浅显，难以得到更深层次的信息。

6.1.2 logistic 模型

优点：

①logistic 模型对变量类型要求不高。因变量为分类变量，自变量可以是离散变量或虚拟变量，也可以是连续变量。如果用线性回归来模拟的话，会产生异方差的现象。

②该模型对数据的要求没有那么严格,不要求原数据的变量满足正态分布,这就客服了线性回归的假设约束，使得适用范围更为广泛。

③可以选择更多的解释变量来增强模型的解释精度，变量的选择范围也更广，并且模型的形式保证了模型的回归结果在有意义的区间值内取值[8]

缺点：

① 对模型中自变量多重共线性较为敏感，例如两个高度相关自变量同时放入模型，可能导致较弱的一个自变量回归符号不符合预期，符号被扭转。需要利用因子分析或者变量聚类分析等手段来选择代表性的自变量，以减少候选变量之间的相关性② 层次分析法是一种带有模拟人脑决策方式的方法，因此带有较多的定性色彩，具有较大的主观性。

②预测结果呈“S”型，因此从 log(odds)向概率转化的过程是非线性的，在两端随着 log(odds)值的变化，概率变化很小，边际值太小，slope 太小，而中间概率的变化很大，很敏感，导致很多区间的变量变化对目标概率的影响没有区分度，无法确定阀值。

6.1.3 层次分析法

优点：

层次分析法具有使用性，将定性与定量相结合，能处理传统优化方法难以解决的问题。且计算简便，结果明确，便于决策者直接了解和掌握。

缺点：

层次分析法是一种带有模拟人脑决策方式的方法，因此带有较多的定性色彩，具有较大的主观性。

6.1.4 主成分分析

优点：

主成分分析利用降维技术用少数几个综合变量来代替原始多个变量，这些综合变量的大部分信息，这些综合变量集中了原来变量的大部分信息，其次它通过计算综合主成分函数得分，对客观经济现象进行科学评价，在次它在应用上侧重于信息贡献影响了力综合评价。

缺点：

当主成分的因子负荷的符号有正有负时，综合评价意义就不明确，命名清晰性低。

6.2 模型的推广

1. 客户画像，即用户信息化标签，通过收集消费者社会属性、生活习惯、消费为等主要信息的数据后，完美的抽象出一个用户的企业全面，可看做是企业应用大数据的基本方式。客户画像为企业提供足够的信息基础，帮助企业快速找到精准用户群体以及用户需求等更为广泛的反馈信息。

2. 层次分析法适用于具有分层交错评价指标的目标系统，而且目标值又难于定量描述的决策问题。层次分析法主要应用在安全科学和环境科学领域。在安全生产科学技术方面主要应用包括煤矿安全研究、危险化学品评价、油库安全评价、城市灾害应急能力研究以及交通安全评价等；除此之外，层次分析法更多的可以用于指导和解决个人生活中遇到的问题，比如说专业的选择、工作的选择以及买房的选择等。

七、 参考文献

[1]杨达，考虑后车的车辆跟驰行为建模分析[D].成都西南交通大学.2013.

[2]白其峥，数学建模案例分析[M].北京：海军出版社,2000.

[3]邓忠，基于车载自组织网络的车辆状态识别与驾驶行为评估[D].广州：州：华南理工大学,2005.
[4]薛定语，陈阳泉，高等应用数学问题的 MATLAB 求解[M].北京：清华出版社,2004.
[5]杨启帆等，数学建模[M].北京：高等教育出版社,2005.
[6]Technimetrics. Logistic Regression: A Self-Learning Text[M]. Springer, 2004.
[7]钟礼杰，高玉堂，金丕焕. logistic 回归模型的拟合优度检验[J] . 中国卫生 统计，1993(3):5 5-59.
[8]张连增，孙维伟. 车险索赔概率影响因素的 Logistic 模型分析[J].保险研究，2012(7): 16-25.

附录

```
A=[]
[n,n]=size(A);
[V,D]=eig(A);
tempNum=D(1,1);
pos=1;
for h=1:n
    if D(h,h)>tempNum
        tempNum=D(h,h);
        pos=h;
    end
end
w=abs(V(:,pos));
w=w/sum(w);
t=D(pos,pos);
disp('准层特征向量 w=');disp(W);disp('准层最大特征根 t=');disp(t);
CI=(t-n)/(n-1);RI=[0 0 0.58 0.90 1.12 1.24 1.32 1.41 1.45 1.49 1.51];
CR=CI/RI(n);
if CR<0.10
    disp('此矩阵的一致性可以接受!');
    disp('CI=');disp(CI);
    disp('CR=');disp(CR);
else disp('此矩阵的一致性检验失败，请重新进行评分!');
end
disp('请输入方案层各因素对准侧层各因素权重的成对比较阵');
for i=1:n
    disp('请输入第');disp(i);disp('个准则层因素的判断矩阵 B');disp(i);
    G=input('=');
    [m,m]=size(G);
    [V,D]=eig(G);
```

```
        tempNum=D(1,1);
        pos=1;
        for h=1:m
            if D(h,h)>tempNum
                tempNum=D(h,h);
                pos=h;
            end
        end
        eval(['W',num2str(i),'=abs(V(:,pos))/sum(abs(V(:,pos)))']);
        eval(['T',num2str(i),'=D(pos,pos)']);
        temp=D(pos,pos);
        CI=(temp-m)/(m-1);RI=[0 0 0.58 0.90 1.12 132 1.41 1.45 1.19 1.51];
        CR=CI/RI(m);
        if CR,0.10
            disp('此矩阵的一致性可以接受!');
        else  disp('此矩阵的一致检验性失败,请重新进行评分并在  clear  后重新运行程序!');return;
        end
        eval (['B',num2str(i),'=G']);
    end
    g=1:n;
    [k,k]=size(B1);
    for i=1:k
        sum=0;
        disp('第');disp(i);disp('个方案的总排序权值为');
        for j=1:n
            change=eval(['W',num2str(j),]);
            sum=sum+change(i,1)*w(j,1);
        end
        eval(['p',num2str(i),'=sum']);
    disp('A(6)')
    A=[ ];%输入判断矩阵
    [~,n]=size(A);%计算 A 的维度
    x=ones(n,100);%x 为 n 行 100 列全为 1 的方阵
    y=ones(n,100);%y 为 n 行 100 列全为 1 的方阵
    m=zeros(1,100);%m 为 1 行 100 列全为 0 的向量
    m(1)=max(x(:,1));%x 第一列的最大值赋值给 m 的第一个向量
    y(:,1)=x(:,1);%x 第一列的赋值给 y 的第一列
    x(:,2)=A*y(:,1);%x 的第二列是 A*y 的第一列得来的
    m(2)=max(x(:,2));%x 第二列的最大值赋值给 m 的第二个分量
```

```
y(:,2)=x(:,2)/m(2);%的第二列除以 m 的第二个分量后赋值给 y 的第二列
p=0.0001;i=2;k=abs(m(2)-m(1));%初始化 p,i,k,k 为 m(2)-m(1)的绝对值
while   k>p%当 k>p 时执行以下循环体
   i=i+1;
   x(:,i)=A*y(:,i-1);%x 的第一列是 A 乘以 y 的第 i-1 列
   m(i)=max(x(:,i));%x 第 i 列的最大值赋值给 m 的第 i 个分量
   y(:,i)=x(:,i)/m(i);%y 的第 i 列是 x 的第 i 列除以 m 的第 i 个分量
   k=abs(m(i)-m(i-1));%k 等于 m(i)-m(i-1)的绝对值
end
a=sum(y(:,i));% y 的第 i 列的和赋值给 a
w=y(:,i)/a;% y 的第 i 列除以 a
t=m(i);% m 的第 i 个分量赋给 t
disp('权向量');disp(w);%  显示权向量 w
disp('最大特征值');disp(t);%  显示最大特征值 t
            %以下是一致性检验
CI=(t-n)/(n-1);% t-维度再除以维度-1 的值赋给 CI
RI=[0 0 0.58 0.90 1.12 1.24 1.32 1.41 1.45 1.49 ];
CR=CI/RI(n);%计算一致性
if CR<0.10
      disp('此矩阵的一致性可以接受!');
      disp('CI=');disp(CI);
      disp('CR=');disp(CR);
else
      disp('此矩阵的一致性不可以接受!');
end
```