1. 问题重述
   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。

可以预见未来车险行业的几大发展趋势：  
1. 车险价格与驾驶行为密切相关  
未来的车险定价将逐渐转变为“从人主义”。车险的定价因素将直接与驾  
驶人的驾驶习惯与行驶里程挂钩，通过驾驶行为来判定车险价格，可能会使  
车险由原来的一年买一次变成可以一个月买一次。一个具有良好驾驶习惯的  
车主，可能只需要支付原本保费的 30% 左右，而驾驶习惯不佳的车主，则会  
在原本保费的基础上继续上涨。  
2. 同价位车型车险价格完全不同  
国内传统的汽车保险定价，通常是以车型和其购置价为主要依据。未来中  
国车险业，同样的一款车，不同的人开，保费价格会完全不同。这个不同可能  
是取决于投保人本身的驾驶行为，还可能会以投保人本身的年龄、职业、家庭  
状况等信息为标准。  
信息时代的到来，为车险企业提供了一个更加有力的武器，可以通过数字  
化技术来更加精准地了解客户，制定营销和服务方案。

* 1. 问题提出
     1. 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。
     2. 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

用户画像是一种全面勾画用户、联系用户与 产品的良好工具，是建立在一系列真  
实数据之上的用户目标模型。用户画像描绘用于标识用户的各项信息，比如用户背景、身份、兴趣、需求、心理、性格、收入、职业等，通过用户画像，我们可以细致地了解到一个用户的信息全貌。随着互联网技术的发展和大数据时代的来临， 在数据驱动下， 用户画像的内涵和外延都发生了变化， 主要是通过数据刻画用户特征， 从而为用户提供优质服务。第一小题要求根据附件一的数据对用户进行精确画像，建立合理的信息模型，给出客户的续保概率，第二小题要求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，好好理解题意，不难发现我们要做的其实是从给定的数据集合中提取出用户的个人信息集合，来独立进行用户需求、偏好和兴趣描述的模型。

提取用户标签是刻画用户画像的关键。应用大数据分析、数据挖掘、机器学习等技术处理和分析用户的行为数据和信息数据是提取用户画像的偏好、兴趣标签的重要途径之一。针对以上提出的两个问题，我们主要采用基于大数据挖掘的用户画像构建方法，首先对附件一中的用户数据进行预处理，主要包括：

* “分类”数据离散化；
* 对附件一中不同维度的表项进行关联性分析，将与其他维度关联性不强且对模型最终预测结果影响很轻微的维度删除；比如附件一中的“品牌”和“车系”维度对最终预测结果影响较小且与其他维度的关联关系很弱，则在提取标签的时候可以不使用这两个维度的数据；
* 对数据进行“归一化”处理，提高模型训练时的收敛速度，提高模型性能；

从预处理后的数据中提取标签，提取的标签应满足重复性低、交叉率低、维度划分明显等基本条件。对数据进行预处理并提取标签之后，就可以建立………