.

**数据挖掘技术结课大作业**

**（2023春季学期）**



|  |  |  |  |
| --- | --- | --- | --- |
| **学生姓名** |  | **学　号** |  |
| **教学院系** | **计算机与科学学院** | | |
| **专业年级** |  | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **完成日期** | **2023** | **年** |  | **月** |  | **日** |

1. 香皂用户的细分。

**1.1 背景分析**

CRISA是一家市场研究机构，专注于跟踪消费者在消费品（包括耐用品和非耐用品）方面的购买行为。在某重要研究项目中，CRISA跟踪了许多消费品类型（如洗涤剂），其中每一种消费品都有几十个品牌。为了跟踪顾客的购买行为，CRISA在某国的100个城镇建立了样本，以覆盖该国绝大部分市场。为了保证样本具有代表性，CRISA使用层次采样法精心挑选这些样本。

数据集中既有交易数据（每行表示一笔交易），也有家庭数据（每行表示一个家庭）。家庭数据包含以下信息：

* 家庭人口的统计信息（每年都会更新）
* 拥有的耐用品（耐用品是指汽车、洗衣机等，每年都会更新可通过这些信息计算富裕指数）
* 所购买商品的类别和品牌等，每个月都会更新。

CRISA有两类客户：一类是广告公司，它们通过订阅数据库服务来获取每个月的最新数据，并利用这些数据向客户提供意见和促销策略；另一类是消费品制造厂家，他们通过CRISA数据库监测市场份额。

* 1. **关键问题**

CRISA以前的做法是根据客户的人口数统计数据细分市场，但是现在，CRISA希望根据两组关键变量细分市场，这两组变量直接与购买过程和品牌忠诚度有关。

* 购买行为（购买数量、购买频率、交易金额和品牌忠诚度）
* 购买基础信息（对折扣的敏感度、价格分类、销售产品分类）

这样做的好处是，CRISA能够获得不同购买行为的人口统计特征和品牌忠诚度等数据，从而帮助客户（在本案例中是一家名为IMRB的公司）更合理地规划促销预算。更有效的市场细分也使得IMRB可以设计多种促销活动—分别针对不同的子市场和一年中的不同时段，而且成本效益更佳。这反过来又促使IMRB把促销预算更合理地分配给不同的子市场，从而制订更有效的激励计划，提高顾客对产品的忠诚度。

* 1. **数据**

下表显示了每个家庭的基本数据（每行代表一个家庭）的说明。

|  |  |  |
| --- | --- | --- |
| 变量类型 | 变量 | 描述 |
| 成员ID | Member id | 每个家庭的唯一标识符 |
| 人口统计信息 | SEC | 社会保险类别（1表示高，5表示低） |
| FEH | 饮食习惯（1表示素食者，2表示素食者但也吃鸡蛋，3表示非素食者，0表示不明确） |
| MT | 母语 |
| SEX | 操持家务者的性别（1表示男性，2表示女性） |
| AGE | 操持家务者的年龄 |
| EDU | 操持家务者的教育程度（1表示最低教育程度，9表示最高教育程度） |
| HS | 家庭成员数量 |
| CHILD | 是否有孩子 |
| CS | 是否有电视（1表示有电视，2表示没有电视） |
| Affluence Index | 所拥有耐用品的带权总价值 |
| 过去一段时间的购买信息摘要 | No. of Brands | 所选购品牌的数量 |
| Brand Runs | 连续购买的品牌数量 |
| Total Volume | 总金额 |
| No. of Trans | 交易次数（如果一个月内购买多个品牌，将被当作不同的交易） |
| Value | 购买金额（意义不是很明确，似乎和Total Volume有关） |
| Trans/Brand Runs | 每个品牌的平均交易次数 |
| Vol/Tran | 每次交易的平均金额 |
| Avg. Price | 每次交易的平均价格 |
| 促销期购买情况 | Pur Vol. No Promo-% | 无促销时交易金额占总金额的百分比 |
| Pur Vol Promo 6% | 在促销规则6下，交易金额占总金额的百分比 |
| Pur Vol Other Promo % | 在其他促销规则下，交易金额占总金额的百分比 |
| 品牌购买情况 | Br. Cd. (57,144),55,272, 286,24,481,352,5, and 999(others) | 相应品牌的交易金额（(57,144)代表一个品牌，其他55、272等一个数字代表一个品牌，others 999代表剩下的其他品牌 |
| 不同价格分类下的交易情况 | Price Cat 1 to 4 | 各个价格分类（1到4类）的交易额占比 |
| 与销售产品分类有关的交易情况 | Proposition Cat 5 to 15 | 各个产品分类下的交易金额占总金额的比例（可以主要考虑那些占比大的产品） |

* 1. 测试品牌的忠诚度

在本案例中，客户购买不同品牌产品的总数代表了客户的忠诚度。例如，假设某客户短期内购买第一和第二个品牌，然后长期购买第三个品牌，那么这位客户的品牌忠诚度肯定不同于不断在这三个品牌间来回切换购买的客户。因此，客户从一个品牌切换到另一个品牌的时间间隔也是衡量品牌忠诚度的重要指标。但是，也有观点认为品牌忠诚度与不同品牌的购买比例有关—一位客户在某品牌上消费了90%的金额，而另一位客户在多个品牌上均等消费，那么前者的品牌忠诚度肯定比后者高。

* 1. 任务

1. 根据以下变量，使用至少两种不同的聚类算法识别相似家庭所属的簇：
2. 描述购买行为的变量（包括品牌忠诚度）
3. 描述购买基础信息的变量
4. 描述购买行为和购买基础信息的变量

（2）评估不同算法、变量得到的聚类效果，选出最优的模型和参数。

提示：如何选择k值？建议首先考虑如何使用簇。市场促销活动支持2-6种不同的促销方式。如何处理不同品牌的购买量占总购买量的比例？只购买品牌A的客户的品牌忠诚度与只购买品牌B的客户的品牌忠诚度一样吗？品牌所占的市场份额对距离度量指标的计算有什么影响？

1. 出租车取消问题

2.1 背景分析

Taxi-cancellation-case.csv包含了从2011年至2013年期间的出租车预定数据，包含约1万条记录，每条记录代表一次预定服务。该数据集包含17个输入变量—用户ID、车型、是通过网站预订还是通过手机应用预订、出行类别、服务类别、地理信息、预订的出行日期和时间等等。目标变量是二元变量，用于表示预订的出租车服务是否会被取消。整体取消率在7%~8%之间。

* 1. 任务

请利用这些数据建立2-4个预测模型，并评估和比较所建预测模型。下面是一些提示：

* 在建立解释性模型时，不必解决所有的数据准备问题。以GPS信息为例，由于还有其他地理信息可供使用，因此建议把解析或使用GPS信息的难题暂时放到一边。
* 如何处理缺失数据？如何处理数据项被标记为NaN的记录？
* 分析日期和时间字段（预订时间戳和行程时间戳）能提供哪些有用信息？
* 考虑如何处理分类变量，需要把它们全都转换为虚拟变量吗？还是只需要转换其中一部分？

备注：结课大作业应以word或者PDF文件提交，包括分析建模，实验结果，模型评估以及总结和结论等。完整源代码作为结课大作业的附件一并提交，关键代码应有清晰、详细的注释说明。