第一阶段：请于3月12日，回复完成小案例解答需要的知识点，简单写出知识点名称即可。如：XX检验、方差、XX算法

第二阶段：请于2018年4月30日之前完成此次小案例分析问题，涉及到软件分析的，请提供分析脚本（软件类型不限），选择题也提供分析的中间过程脚本语言或图。

根据相同的背景材料和数据回答以下的题目。题目分为两个部分。第一部分为选择题，每道题的答案个数不固定。在做题过程中需要使用统计软件进行相应的操作。提供CSV格式的数据，统计软件不受限制。各题A)、B)、C)、D)四个选项中，每题至少有一个选项是正确的。第二部分为模型实作题，使用与第一部分一样的数据，但最终须提交对测试数据的预测结果。I、某电信公司想针对用户手机通话记录，希望从用户历史通话数据中，运用数据挖掘(Data Mining)中的分类(Classification)技术，分析出会影响电信用户流失的相关因素，并建立能预测用户是否流失的数据挖掘分类模型。

|  |  |
| --- | --- |
| 字段名称 | 标签 |
| subcriberID | 用户编号 |
| Churn | 是否流失（1=流失） |
| Gender | 性别（1=男） |
| AGE | 年龄 |
| edu\_class | 教育等级：0=小学及以下，1=中学，2=本科，3=研究生 |
| incomeCode | 用户居住区域平均收入代码 |
| Duration | 在网时长（月） |
| Feton | 是否飞信用户（1=开通） |
| peakMinAv | 最高单月通话时长 |
| peakMinDiff | 统计期内结束月份与开始月份通话时长增长量 |
| posTrend | 用户通话是否有上升态势（1=是） |
| negTrend | 用户通话是否有下降态势（1=是） |
| nrProm | 营销次数 |
| Prom | 最近一个月是否被营销（1=是） |
| curPlan | 统计开始时套餐类型（1=200分钟，2=300分钟，3=350分钟，4=500分钟 |
| avgPlan | 统计期内使用时间最长的套餐 |
| planChange | 更换套餐（-3至3的整数， 代表降低或提升套餐的档位数量） |
| posPlanChange | 统计期内是否提高套餐（1=是） |
| negPlanChange | 统计期内是否降低套餐（1=是） |
| Call\_10086 | 是否拨打过客服电话（1=是） |

1、训练集当中哪个字段有缺失值（D）

A.prom

B.incomeCode

C.以上皆是

D.以上皆非

2、训练集中受教育程度edu\_class的众数是（B）

A.0

B.1

C.2

D.3

3、训练集中关于男性和女性的离网率说法正确的是（A、C）

A.男性离网率比较高

B.女性离网率比较高

C.男性离网率约为52.8%

D.女性离网率约为52.8%

4、在训练集当中筛选变量时，在选定的离散变量（gender, edu\_class, feton, prom, posPlanChange, curPlan, call\_10086）中，使用卡方检验选择与用户是否离网（churn）相关性较大的变量，设定显著性水平为0.05，那么需要删除的变量是（A、B）

A.prom

B.posPlanChange

C.curPlan

D.call\_10086

5、在应用列联表分析feton与call\_10086之间的关系时，可以得到它们都是“1”的期望频率为（C）

A. 457

B.609

C.804

D.522

6、训练集中为了改善duration的偏态，我们对其取自然对数后命名为log\_duration，变换后，考虑用四分位距判断log\_duraion是否存在离群值，以下关于log\_duration正确的是（B、D）

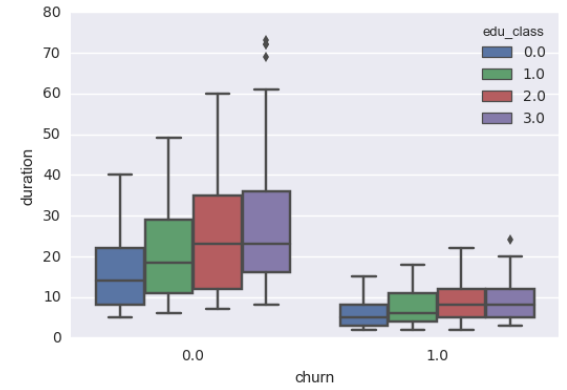
A.存在离群值

B.不存在离群值

C.四分位距约为1.7

D.四分位距约为1.2

7、在进行数据探索时，经常会使用图形以直观地发现数据表现出的规律，下图是关于教育程度、在网时长、是否离网三个变量的关系的图形，就此图形来看的话，说法正确的是（A、C）



A.离网用户的在网时长更短

B.教育程度较低的用户越容易离网

C.教育程度低的用户相对在网时长会更短

D.图中箱线图的箱体越长（四分位距越大），则方差一定越大（观察3.0与2.0可以知道此选项是错的）

8、训练集中对duration进行离散化，关于以下说法正确的是（C）

A.等宽离散化后每箱的样本量会大致相等（等宽的K值是多少？）

B.处理极端值要在离散化之前进行（之后）

C.人工进行离散化时，在(4, 8]一段的用户离网率约60%

D.无论如何进行离散化，离网率都是随着duration增加而增加（不是，是增加而减少）

9、训练集中planChange代表用户更换套餐的档次，取值是从-3到3之间的整数，负值代表降低套餐，正值代表提升套餐，0代表未改变过套餐，数字大小代表了用户降低后提升套餐的档位数，例如-2代表了用户将自己的套餐降低了两个档位。通过分析，判断以下结论正确的是（A、C）

A.planChange不同分类水平下，用户的离网率有较大差异，因此建模时应当保留

B.planChange不同分类水平下，用户的数量有较大差异，因此建模时应当保留

C.planChange与churn变量的卡方检验显示两者是相关的，因此建模时应当保留

D.以上皆非

10.对于本例中的预测来说，我们更关注流失用户的发现，为此生成两个模型，混淆矩阵分别为model1和model2，那么对于比较两个模型的效果，以下结论正确的是（B、D）

model1

| **prediction** | **0.0** | **1.0** |
| --- | --- | --- |
| **churn** |  |  |
| **0.0** | 1286 | 60 |
| **1.0** | 432 | 646 |

model2

| **prediction** | **0.0** | **1.0** |
| --- | --- | --- |
| **churn** |  |  |
| **0.0** | 1264 | 82 |
| **1.0** | 294 | 784 |

A.model1的灵敏度recall大于model2

B.model1的命中率precision大于model2

C.model1的F-measure大于model2

D.效果上来看，model2更好

II、请以训练数据telecom\_train.csv为基础，建立一个预测客户是否会流失的分类预测模型。利用此模型对测试数据telecom\_test.csv中的数据进行预测，并输出一个测试结果的档案(results.csv)。results.csv中只有两个字段，分别是ID及预测是否流失(1(流失)/0(未流失)) 。results.csv的形式如下：

| **subscriberID** | **prediction** |
| --- | --- |
| 74072181 | 0 |
| 74929937 | 0 |
| 74923700 | 0 |
| 72235429 | 0 |

模型的评估方式是以流失的F-measure为主。

朴素贝叶斯分类器：F1 = 0.7416812609457092

神经网络分类器： F1 = 0.792303897385