|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 变量名 | | 详细说明 | 备注 |
| 因变量（下月） | churn | 是否流失 | 1=流失  0=不流失 | 流失率1.27% |
| 自变量（当月） | tenure | 在网时长 | 连续变量  单位：天 | 客户从入网到截止数据提取日期时在网时间 |
| expense | 当月花费 | 连续变量  单位：元 | 客户在提取月份时的花费总额 |
| degree | 个体的度 | 连续变量  单位：人数 | 和客户通话的总人数，去重之后的呼入与呼出加总 |
| tightness | 联系强度 | 连续变量  分钟/人 | 通话总时间除以总人数 |
| entropy | 个体信息熵 | 连续变量 | ，其中为个体i的信息熵，代表个体i和j通过电话，代表j和i通话的分钟数据占i总通话分钟的比例 |
| chgdegree | 个体度的变化 | 连续变量  单位：% | （本月个体的度-上月个体的度）/上月个体的度 |
| chgexpense | 花费的变化 | 连续变量  单位：% | （本月花费-上月花费）/上月花费 |

**数据说明：**

* **训练数据集：sampledata.csv**
* **测试数据集：predata.csv**

数据文件来自国内某运营商，数据已经进行了清理，数据集中的变量包括：是否流失（churn）、在网时长（tenure）、当月花费（expense）、个体的度（degree）、联系强度（tightness）、个体信息熵（entropy）、个体度的变化（chgdegree）、花费的变化（chgexpense）共8个变量。具体的变量说明表如下所示：

**分析任务：**

1. 读入训练数据，删除变量expense、tightness、degree中的负值，完成数据清洗；

2. 绘制因变量和各个自变量的箱线图（提示：可以对右偏分布的数据取对数）；

3. 以是否流失为因变量，对自变量进行标准化（使其均值为0，方差为1，提示：在R中可使用scale函数），建立逻辑回归模型，给出系数估计结果，并对结果进行解读（提示：使用glm()函数建立逻辑回归模型）；

4. 使用建立好的逻辑回归模型，分别对**训练集和测试集**进行预测，得到每个用户的预测流失概率值（提示：使用predict()函数进行预测）；

5. 借助问题4中预测的结果，分别绘制训练集和测试集上预测结果的ROC曲线，计算相应的AUC值，并根据ROC曲线和AUC值对模型进行评价（提示：使用R包pROC中的plot.roc()函数绘制ROC曲线）。