实验一：通信客户流失预警模型

## 1挖掘过程

#### 1.1业务理解

使用分类模型构建客户流失预测模型，通过客户为流失客户的概率预测该客户是否为流失客户并根据客户为流失客户的概率生成流失概率排序名单。

#### 1.2数据理解

在本案例中使用的数据来自于IBM Sample Data Sets，是某电信公司一段时间内的客户消费数据。共包含7043笔客户资料，每笔客户资料包含21个字段，其中1个客户ID字段，19个属性特征字段及1个标签字段（Yes代表流失，No代表未流失）。属性特征字段主要包含以下三个维度指标：客户画像指标（如性别、是否老年等）、消费产品指标（如是否开通互联网服务、是否开通电话服务等）、消费信息指标（如付款方式、月费用等）。字段的具体说明如表 1所示：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **字段** | **字段翻译** | **角色** | **特征类型** | **不同值个数** | **备注** |
| **customerID** | 客户ID | ID | 无类型 | 7043 |  |
| **Gender** | 性别 | 特征 | 类别 | 2 |  |
| **SeniorCitizen** | 是否为老年人 | 特征 | 类别 | 2 | 初始取值为0/1，分别代表否/是 |
| **Partener** | 是否有配偶 | 特征 | 类别 | 2 |  |
| **Dependents** | 是否和需抚养/赡养人同居 | 特征 | 类别 | 2 |  |
| **Tenure** | 在网时长 | 特征 | 数值 |  | 单位：月 |
| **PhoneService** | 是否开通电话服务业务 | 特征 | 类别 | 2 |  |
| **MultipleLines** | 是否开通多线业务 | 特征 | 类别 | 3 | 取值为是/否/未开通电话业务 |
| **InternetService** | 是否开通互联网服务 | 特征 | 类别 | 3 | 取值为否/DSL/光纤 |
| **OnlineSecurity** | 是否开通网络安全服务 | 特征 | 类别 | 3 | 取值为是/否/未开通互联网服务 |
| **OnlineBackup** | 是否开通在线备份业务 | 特征 | 类别 | 3 | 取值为是/否/未开通互联网服务 |
| **DeviceProtection** | 是否开通了设备保护业务 | 特征 | 类别 | 3 | 取值为是/否/未开通互联网服务 |
| **TechSupport** | 是否开通了技术支持服务 | 特征 | 类别 | 3 | 取值为是/否/未开通互联网服务 |
| **StreamingTV** | 是否开通网络电视 | 特征 | 类别 | 3 | 取值为是/否/未开通互联网服务 |
| **StreamingMovies** | 是否开通网络电影 | 特征 | 类别 | 3 | 取值为是/否/未开通互联网服务 |
| **Contract** | 签订合同方式 | 特征 | 类别 | 3 |  |
| **PaperlessBillling** | 是否开通电子账单 | 特征 | 类别 | 2 |  |
| **PaymentMethod** | 付款方式 | 特征 | 类别 | 4 |  |
| **MonthlyCharges** | 月费用 | 特征 | 数值 |  | 单位：美元 |
| **TotalCharges** | 总费用 | 特征 | 数值 |  | 单位：美元 |
| **Churn** | 是否流失 | 标签类别 | 类别 | 2 | 取值为是/否 |

表 1 数据集字段说明

#### 1.3数据准备

本案例数据准备工作步骤如下。

##### 读入数据集

代码如表 2所示：

|  |
| --- |
| 读入数据集代码 |
| import numpy as np |
| import pandas as pd |
|  |
| df = pd.read\_csv('./Telco-Customer-Churn.csv') |
| df.head() |

表 2 读入数据集代码

##### 数据初步清洗

首先进行初步的数据清洗工作，包含错误值和异常值处理，并识别字段类型。其中清洗工作主要包含：

1. 对MultipleLines（是否开通多线服务）、OnlineSecurity（是否开通网络安全服务）、OnlineBackup（是否开通在线备份业务）、DeviceProtection（是否开通了设备保护业务）、TechSupport（是否开通了技术支持服务）、StreamingTV（是否开通网络电视）、StreamingMovies（是否开通网络电影）等属性特征进行错误值处理。如不满足前置条件，则这些特征直接取值为No（若未开通互联网服务，自然不会开通网络电视等服务）。
2. 对TotalCharges（总费用）特征进行异常值处理。将空白取值替换空值nan，由于含空值的行较少，将其一并删除。删除后该特征取值只有数值，因此将此特征类型设为浮点型。
3. 对Tenure（在网时长）属性离散化。Tenure取值为[0, 72]，为适应后续决策树建模，将其离散化，以12、24、36、48、60为临界值进行分箱操作，共分为6组。
4. 判断字段的类别

数据清洗代码如表 3所示：

|  |
| --- |
| 数据初步清洗代码 |
| # 错误值处理 |
| repl\_columns = ['OnlineSecurity', 'OnlineBackup', 'DeviceProtection', |
| 'TechSupport','StreamingTV', 'StreamingMovies'] |
|  |
| for i in repl\_columns: |
| df[i] = df[i].replace({'No internet service' : 'No'}) |
|  |
| # 替换值SeniorCitizen，便于后续探索性分析 |
| df["SeniorCitizen"] = df["SeniorCitizen"].replace({1: "Yes", 0: "No"}) |
|  |
| # 替换值TotalCharges |
| df['TotalCharges'] = df['TotalCharges'].replace(' ', np.nan) |
|  |
| # TotalCharges空值：数据量小，直接删除 |
| df = df.dropna(subset=['TotalCharges']) |
| df.reset\_index(drop=True, inplace=True) # 重置索引 |
|  |
| # 转换数据类型 |
| df['TotalCharges'] = df['TotalCharges'].astype('float') |
|  |
| # 转换tenure |
| def transform\_tenure(x): |
| if x <= 12: |
| return 'Tenure\_1' |
| elif x <= 24: |
| return 'Tenure\_2' |
| elif x <= 36: |
| return 'Tenure\_3' |
| elif x <= 48: |
| return 'Tenure\_4' |
| elif x <= 60: |
| return 'Tenure\_5' |
| else: |
| return('Tenure\_over\_5') |
|  |
| df['tenure\_group'] = df.tenure.apply(transform\_tenure) |
|  |
| # 数值型和类别型字段 |
| Id\_col = ['customerID'] |
| target\_col = ['Churn'] |
|  |
| cat\_cols = df.nunique()[df.nunique() < 10].index.tolist() |
| num\_cols = [i for i in df.columns if i not in cat\_cols + Id\_col] |
|  |
| print('类别型字段：\n', cat\_cols) |
| print('-' \* 30) |
| print('数值型字段：\n', num\_cols) |

表 3 数据初步清洗代码

输出结果如图 1所示：

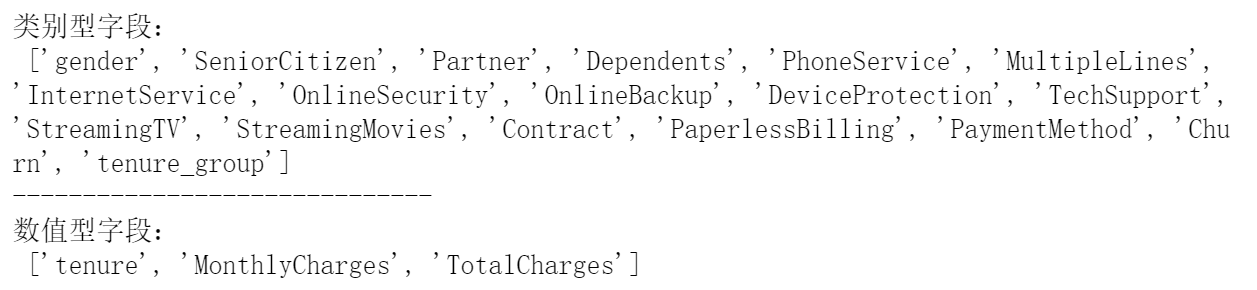


图 1 类别型字段与数值型字段

##### 数据初步清洗

##### 数据预处理

。。。。。。。。。。。。。。。

#### 1.4模型构建

。。。。。。。。。。。。。

## 实验小结

主要工作小结

通过案例数据挖掘得到的成果、对客户保留计划的建议

本实验过程中的创新点