# xx公司现金贷客户信用评分卡报告

模型开发开始日期：2017-12-01

模型开发结束日期：2017-12-02

上传报告日期：2017-12-03

责任人：匡昊

## 1数据基本情况

描述数据基本情况

### 1.1数据集情况

#### 1.1.1数据集大小

建模数据样本共xx行，变量个数x,目标变量:has\_overdue

#### 1.1.2变量名称&类型

模型数据变量列表

|  |  |  |  |
| --- | --- | --- | --- |
| **字段** | **名称** | **类型** | **单位** |
| sex | 性别 | 分类 |  |
| Contact\_sex | 联系人性别 | 分类 |  |
| contact\_age | 联系人年龄 | 连续 | 岁 |
| age | 年龄 | 连续 | 岁 |
| duration\_days | 客户在网时长 | 连续 | 月 |
| recently\_flow\_pos1 | 近1个月交易流水 | 连续 | 金额（元） |
| recently\_flow\_pos2 | 近2个月交易流水 | 连续 | 金额（元） |
| recently\_flow\_pos3 | 近3个月交易流水 | 连续 | 金额（元） |
| …………………………………… | | | |
| Has\_overdue | 是否逾期 | 分类 |  |

### 1.2缺失值

#### 1.2.1缺失值情况

1.各变量缺失值情况

|  |  |
| --- | --- |
| **变量** | **缺失值个数** |
| sex | 3 |
| Contact\_sex | 3 |
| avg\_price\_trade\_pos1 | 8 |
| avg\_price\_trade\_pos2 | 21 |
| avg\_price\_trade\_pos3 | 29 |
| avg\_price\_trade\_scan1 | 64 |
| avg\_price\_trade\_scan2 | 52 |
| avg\_price\_trade\_scan3 | 53 |

#### 1.2.2缺失值处理

1.删除

行删除方法：对缺失比例超过50%的样本进行删除，共删除xx条样本

列（变量）删除：对缺失比例超过30%的变量进行删除，共删除1个变量

2.填充情况

|  |  |  |
| --- | --- | --- |
| **变量名** | **填充方法** | **填充值** |
| AGE | 均值填充 | 32 |
| SEX | 众数填充 | 0 |

………………….

### 1.3异常值

#### 1.3.1异常值情况

异常值情况表

|  |  |  |  |
| --- | --- | --- | --- |
| **变量** | **异常值示例** | **成因** | **处理** |
| duration\_days | 6500 | 数据库记录错误 | 均值修正 |

#### 1.3.2异常值处理汇总

通过对异常值进行处理，共删除0条记录、修正1条记录

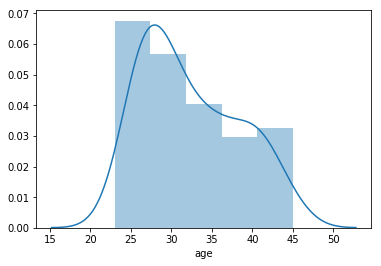
### 1.4单变量统计

对单个变量进行统计分析，以观察数据集各个变量的分布情况，分析数据特征。具体分析详情如下。

#### 1.4.1连续型变量

变量age基本情况分析

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 25% | 50% | 75% | max |
| AGE | 30000 | 32.3452 | 5.94 | 23 | 27 | 31 | 37.25 | 45 |



从以上信息上看，客户年龄主要分布在30岁上下，年级最大的有63岁，最小的有17岁，客户主要年龄段为中青年

……………………………………………………..

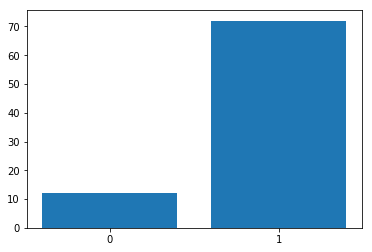
#### 1.4.2分类型变量

此处统计分类型变量，一共xx个变量

变量：sex, contact\_sex…….

方法：条形图&频数统计

性别分布图（1-男，0-女）



性别频数统计结果

|  |  |
| --- | --- |
| 用户性别分布 |  |
| 0 | 305 |
| 1 | 2735 |

从以上图表上看，用户主要分布在男性（1）上…….

………………….

### 1.5多变量统计

汇总多变量统计分析结果，以发现变量与目标变量关联程度，已发现各个变量对客户是否违约的预测能力。

#### 1.5.1相关系数

相关系数矩阵

#### 1.5.2信息值

信息值列表

|  |  |
| --- | --- |
| **变量** | **Info value** |
| sex | 0.009035 |
| contact\_sex | 0.187801 |

……………………..

备注：info value在0.02以下说明变量对目标变量缺乏预测力，0.3以上说明预测力很强

分析：从上表可以看出，变量sex预测力较弱，而contact\_sex…有一定预测能力

## 2模型

### 2.1模型简介

#### 2.1.1模型情况

变量选择方法：向后逐步回归，人工选择

变量预处理：

根据以上统计结果，对各个变量进行数据预处理，方法和相应变量如下：

变量预处理方法与变量对应表

|  |  |
| --- | --- |
| 最优分箱 | special\_trade\_times，… |
| 最优降低基数 |  |
| woe转换 | contact\_sex，….. |
| log转换 | recently\_flow\_pos2 |
| 变量删除 | overdue\_amount |
| Onehot编码 | scan\_flow1 |

变量个数：23

训练集大小：xx行

测试集大小：xx行

结果：通过变量选择方法，共xx个变量进入模型

### 2.2参数估计和拟合评价

#### 2.2.1参数估计

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **变量** | **参数** | **标准误** | **Wald卡方** | **P值** | **95%下界** | **95%上界** |
| const | 22.2762 | 10.001 | 2.227 | 0.026 | 2.674 | 41.878 |
| distance\_apply\_days | 1.2972 | 0.687 | 1.889 | 0.059 | -0.049 | 2.643 |
| recently\_flow\_pos2 | -0.3560 | 0.201 | -1.774 | 0.076 | -0.749 | 0.037 |
| apply\_times | -10.0099 | 4.478 | -2.235 | 0.025 | -18.787 | -1.233 |
| scan\_1 | 6.0478 | 2.829 | 2.138 | 0.033 | 0.504 | 11.592 |
| special\_trade\_times | 5.4152 | 2.489 | 2.176 | 0.030 | 0.537 | 10.294 |

#### 2.2.2拟合评价

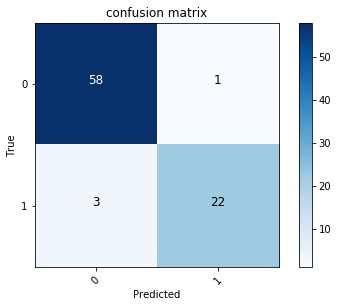
AIC: 28.8726

BIC: 43.4575

分析：模型整体参数估计结果比较显著

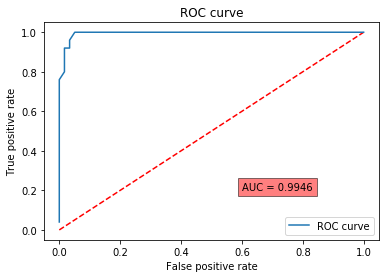
### 2.3模型评估

#### 2.3.1混淆矩阵分析



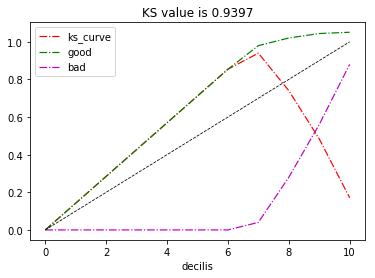
分析:从上图看,模型灵敏度precison=0.96，召回率recall=0.88，模型的预测能力较好

#### 2.3.2 ROC曲线与AUC值



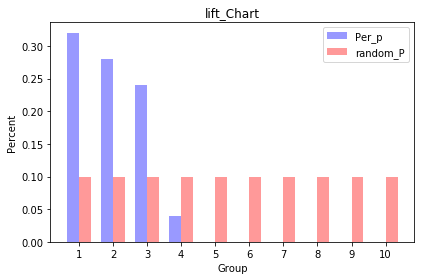
分析：AUC值=0.9946，同时，从ROC曲线面积上看，模型的预测能力非常好

#### 2.3.3 ks曲线与ks值

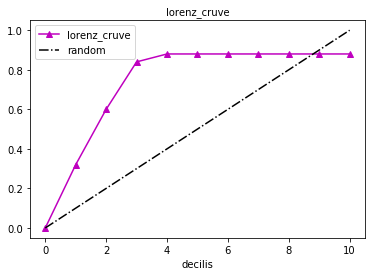


分析：KS值=0.9397,模型对好坏客户的区分能力很强

#### 2.3.4提升图&洛伦茨曲线



分析：从提升图上看，模型预测结果比较稳定。



分析：洛伦茨曲线较陡，模型的预测能力不错。

## 3总结

综上分析，通过本次数据建立的logistic模型拟合情况较好，具备较强的预测能力，可以通过该模型生成评分卡，相应评分卡详情见界面: