申请风险模型开发文档

(加贷后标签)

编写人：方刚

日期：2017/7/10

目录

[1. 项目名称 3](#_Toc489893934)

[2. 项目开发者 3](#_Toc489893935)

[3. 项目目标和基础数据介绍 3](#_Toc489893936)

[3.1 项目目标 3](#_Toc489893937)

[3.2 基础数据介绍 3](#_Toc489893938)

[3.2.1 信审系统主要数据 3](#_Toc489893939)

[3.2.2 信审系统基本流程图 4](#_Toc489893940)

[3.2.3 本次项目开发用到的数据表 4](#_Toc489893941)

[4. 模型标签、特征及特征抽取方案介绍 4](#_Toc489893942)

[4.1 申请风险模型及其标签介绍 4](#_Toc489893943)

[4.2 模型特征介绍 5](#_Toc489893944)

[4.3 模型特征处理 9](#_Toc489893945)

[5. 模型实验结果 12](#_Toc489893946)

[5.1 项目应用的模型 12](#_Toc489893947)

# 1. 项目名称

名称：申请风险模型

# 2. 项目开发者

郑大念，蔡建明，方刚

# 3. 项目目标和基础数据介绍

## 3.1 项目目标

申请风险模型是替代传统的机器审核过程，对于申请信用贷款的用户，通过申请风险模型给出贷款或者拒绝贷款的决定。

## 3.2 基础数据介绍

申请风险系统中目前有信审2.0和信审3.0两套系统，其中信审3.0只包含3.0上线之后的数据，但是这两套系统的数据是同步的。

### 3.2.1 信审系统主要数据

目前信审2.0主要包括以下数据产品：薪新贷、精英贷**、**新薪宜楼贷、精英贷（银行合作）、新薪贷（银行合作）。信审3.0包括薪新贷、助业宜楼贷、精英贷、助业贷、新薪宜楼贷、精英贷（银行合作）、精英贷（农贷）、线下金卡贷0.78-B（综）、线下信用卡189-B（综）、新薪贷（银行合作）、车主现金贷数据产品。

表1是信审系统中涉及的主要数据表信息：

**表1**

|  |  |  |
| --- | --- | --- |
| 表名 | 中文名 | 描述 |
| T\_C\_ALL\_TRANSPORT | 客户进件表 | 记录信审3.0系统所有客户的进件信息 |
| TC\_BS\_TRANSPORT | 客户进件表 | 记录信审2.0系统所有客户的进件信息 |
| TC\_BUSI\_DECISION\_REFUSE | 信审拒绝表 | 记录信审拒绝相关的信息,包括进件的id,拒绝码、拒绝的原因等 |
| ods.clic\_TC\_BS\_TRANSPORT | 信贷进件表 | 记录信贷客户进行相关信息 |
| ods.clic\_TC\_BUSI\_DECISION | 信贷决策表 | 记录信审决策相关的信息 |
| ods.clic\_TC\_BUSI\_DECISION\_REFUSE | 信审拒绝表 | 记录信审拒绝相关的信息,包括进件的id,拒绝码、拒绝的原因等 |
| dw.fact\_tcsv\_contract | 贷后合同表 | 记录客户贷后合同相关的信息，包括合同的开始和结束日期等 |
| dw.fact\_tcsv\_repayment\_plan | 还款计划表 | 记录客户贷后还款信息，包括每期逾期金额，每期应还日期、每期实际还款日期等 |

### 3.2.2 信审系统基本流程图

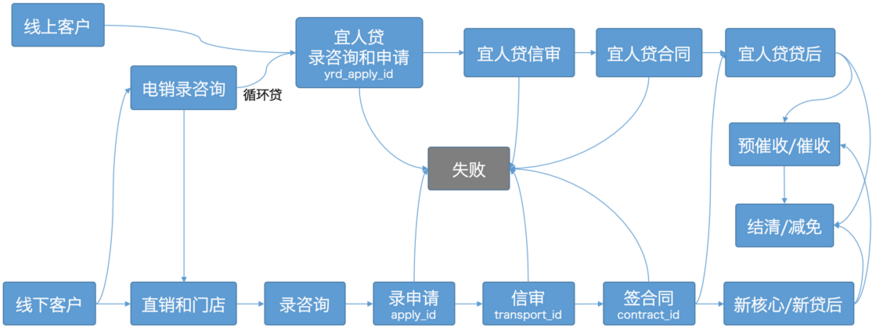


图1 信审基本流程图

### 3.2.3 本次项目开发用到的数据表

**表2**

|  |  |
| --- | --- |
| 表名 | 描述 |
| TC\_MORTGAGOR | 借款人基本信息相关表 |
| TC\_PBOC\_STATISTICS\_INFO | 人行征信信息相关表 |
| TC\_PBOC\_LOAN\_SUMMARY | 借贷信息汇总表 |
| TC\_PBOC\_LOAN\_DETAIL | 借贷信息详细表 |
| TC\_LCSP\_TITLE | 安融征信信息相关表 |
| TC\_LCSP\_QUERY\_DETAIL | 安融征信查询详细表 |
| TC\_MORTGAGOR\_SALARY | 借款人薪资信息相关表 |
| TC\_PBOC\_WELFARE | 借款人福利信息相关表 |
| TC\_CREDIT\_GRADE | 信用评分等级表 |
| TC\_ANTI\_SCORE | 评分信息相关表 |
| TC\_BS\_TRANSPORT | 借款人申请信息相关表 |
| TC\_LOAN\_PURPOSE | 借贷目的相关表 |
| T\_BEE\_TRANSPORT | 进件信息相关表 |
| TC\_HISTORYGUIHU\_DETAIL | 宜信历史信息 |

# 4. 模型标签、特征及特征抽取方案介绍

## 4.1 申请风险模型及其标签介绍

表现变量：用户进行信用申请后，信用审核的结果，即审核通过或者不通过：

坏客户，机审、人工审核阶段不通过、终审和复审阶段不通过的以及贷后违约的（即A、B、C、D）

好客户，终审和复审阶段信用申请通过且贷后未发生相应违约情况(即E)

欺诈用户以及信息虚假用户：这部分数据应排除在模型之外

注：贷后违约指的是在贷后的6期观察期内存在连续3期及其以上的逾期情况

训练集采用时间段在2016-08-01～2017-02-28内的客户数据，测试集采用时间段在2017-03-01~2017-03-09内的数据

**贷后**

放款

违约

**D**

**训练数据**

未违约

**E**

通过

通过

**终审、复审**

**人审**

**机审**

不通过

不通过

不放款

**C**

**B**

**A**

图2 申请风险模型标签

## 4.2 模型特征介绍

特征共分成六个部分：基本信息、征信信息、收入信息、评分信息、申请信息以及宜信历史信息，总计160个特征

**基本信息特征：40个**

|  |  |
| --- | --- |
| 字段名称 | 中文名称 |
| ID\_TYPE | 证件类型 |
| ID\_NUMBER | 证件号码 |
| ID\_VALIDITY\_DATE | 证件有效期 |
| GENDER | 性别 |
| age | 客户年龄 |
| DOMICILE\_PROVINCE | 户籍所在省 |
| domicile\_city | 户籍所在市 |
| RESIDENT\_PROVINCE | 家庭住址所在省 |
| resident\_city | 家庭住址所在市 |
| IDENTITY\_ADDRESS\_PROVINCE | 身份证详细地址省 |
| identity\_address\_city | 身份证详细地址市 |
| mobile\_type | 电话类型 |
| ORG\_TYPE | 单位性质 |
| ORG\_PROVICE | 单位所在省 |
| RECRUITMENT\_DATE | 在单位工作时间 |
| JOB\_POSITION | 职务 |
| MAX\_DIPLOMA | 最高学历 |
| MARRIAGE | 婚姻状况 |
| HAS\_CHILDREN | 有无子女 |
| HOUSE | 房产情况 |
| MONTH\_PAY | 月供 |
| HOUSE\_CITY\_REAL | 是否是本市房产 |
| MONTHLY\_OUTLAY | 月生活费 |
| a\_ANNUAL\_INCOME | 个人年收入 |
| HAS\_CAR | 车产情况 |
| HOUSE\_CONDITION | 居住情况 |
| IN\_CITY\_YEARS | 在此生活年数 |
| BORROWER\_TYPE | 借款人类型 |
| CUSTOMER\_IDENT\_TYPE | 客户身份 |
| SHARE\_PROPORTION | 借款人产权占比 |
| CORP\_REGIST\_CAPITAL | 企业注册资本 |
| PROVIDE\_FOR\_COUNT | 供养人数 |
| INDUSTRY1 | 所属行业 |
| contact\_num | 联系人数量 |
| PLEDGE\_num | 抵押不动产数量 |
| IS\_ABNORMAL | 稽查是否异常 |
| VERIFYING\_RESULT | 联系人调查结果 |
| fam\_stable | 家庭稳定性 |
| risk\_industry | 是否风险行业 |
| IS\_CITY\_SAME | 进件城市与居住城市是否一致 |

**征信相关特征：78个**

|  |  |
| --- | --- |
| 字段名称 | 中文名称 |
| HAS\_CREDIT\_EVENT | 是否具有信用交易信息 |
| CREDIT\_CARD\_NUM | 贷记卡张数 |
| NOMAL\_CARD\_NUM | 正常贷记卡张数 |
| CLOSED\_CARD\_NUM | 销户贷记卡张数 |
| CANCELED\_CARD\_NUM | 注销贷记卡张数 |
| NON\_ACTIVATE\_CARD\_NUM | 未激活贷记卡张数 |
| STOP\_PAY\_CARD\_NUM | 止付贷记卡张数 |
| BLOCKED\_CARD\_NUM | 冻结贷记卡张数 |
| BAD\_DEPT\_CARD\_NUM | 呆账贷记卡张数 |
| CREDIT\_CARD\_MAX\_NUM | 信用卡最大开卡月数 |
| LOAN\_MAX\_NUM | 贷款最大使用月数 |
| MAX\_CREDIT\_CARD\_AGE | 贷记卡最大账龄 |
| MAX\_CREDIT\_LINE | 贷记卡最大授信额度 |
| BLOCKED\_NUM | 贷记卡近12个月内“冻结”“呆账”个数 |
| TOTAL\_CARD\_OVERDUE\_AMOUNT | 贷记卡逾期总额 |
| G\_NUM | 贷记卡出现G的个数 |
| GDZ\_NUM | 贷款出现G、D或者Z的个数 |
| OVERDUE\_NUM\_C1 | 贷记卡最近3个月最高逾期数 |
| OVERDUE\_NUM\_C2 | 贷记卡最近4-6个月最高逾期数 |
| OVERDUE\_NUM\_C3 | 贷记卡最近7-12个月最高逾期数 |
| OVERDUE\_NUM\_C4 | 贷记卡最近13-24个月最高逾期数 |
| OVERDUE\_NUM\_C5 | 贷记卡最近24个月最高逾期数 |
| OVERDUE\_NUM\_C6 | 贷记卡最近12个月最高逾期数 |
| TOTAL\_OVERDUE\_NUM\_C | 贷记卡当前最高逾期期数 |
| MAX\_LOAN\_AGE | 贷款最大账龄 |
| TOTAL\_LOAN\_OVERDUE\_AMOUNT | 贷款逾期总额 |
| HAS\_LOAN\_DUE | 是否存在一个月内到期的贷款 |
| OVERDUE\_NUM\_L1 | 贷款最近3个月最高逾期期数 |
| OVERDUE\_NUM\_L2 | 贷款最近4-6个月最高逾期期数 |
| OVERDUE\_NUM\_L3 | 贷款最近7-12个月最高逾期期数 |
| OVERDUE\_NUM\_L4 | 贷款最近13-24个月最高逾期期数 |
| OVERDUE\_NUM\_L5 | 贷款最近24个月最高逾期期数 |
| OVERDUE\_NUM\_L6 | 贷款最近12个月最高逾期期数 |
| TOTAL\_OVERDUE\_NUM\_L | 贷款当前最高逾期期数 |
| TOTAL\_LOAN\_AMOUNT | 贷款总金额 |
| QUERY\_TIMES1 | 近3个月查询次数 |
| QUERY\_TIMES2 | 近6个月查询次数 |
| OVERDUE\_CARD\_NUM | 发生过逾期的贷记卡张数 |
| MAX\_OVERDUE\_NUM\_C | 贷记卡最大逾期次数 |
| MAX\_OVERDUE\_NUM\_L | 贷款最大逾期次数 |
| OVERDUE\_CARD\_NUM90\_DAYS | 贷记卡最大90天以上逾期次数 |
| BLOCKED\_COUNT | 贷记卡“冻结”“呆账”个数 |
| TOTAL\_OVERDUE\_COUNT\_L | 发生过逾期的贷款个数 |
| OVERDUE\_LOAN\_NUM90\_DAYS | 贷款最大90天以上逾期次数 |
| LOAN\_DUE\_NUM | 一个月内到期的贷款笔数 |
| HOUSING\_LOAN\_NUM | 个人住房贷款笔数 |
| NOMAL\_HOUSING\_LOAN\_NUM | 正常个人住房贷款笔数 |
| QUERY\_TIMES3 | 近24个月以信用卡办理为理由的查询次数 |
| QUERY\_TIMES4 | 近12个月以信用卡办理为理由的查询次数 |
| QUERY\_TIMES5 | 近6个月以信用卡办理为理由的查询次数 |
| QUERY\_TIMES6 | 近24个月以贷款申请为理由的查询次数 |
| QUERY\_TIMES7 | 近12个月以贷款申请为理由的查询次数 |
| QUERY\_TIMES8 | 近6个月以贷款申请为理由的查询次数 |
| QUERY\_TIMES9 | 近12个月查询次数 |
| NOMAL\_CARD\_TOTAL\_AMOUNT | 正常贷记卡总额度 |
| NOMAL\_CARD\_USED\_AMOUNT | 正常贷记卡已使用额度 |
| OUTSTANDING\_LOAN\_REMAIN | 未结清贷款的贷款余额 |
| OUTSTANDING\_LOAN\_CONTRACT | 未结清贷款的贷款合同金额 |
| MAX\_LOAN\_AMOUNT | 贷款最大额度 |
| SUM\_CHARGE\_AGAINST\_LIABILITIES | 总计入负债金额，信用卡汇总信息 |
| SUM\_OVERDRAW\_180 | 总透支180天以上未付余额，贷记卡/信用卡汇总信息 |
| SUM\_CURRENT\_MONTH\_BILL\_AMOUNT | 本月总应还款金额 |
| SUM\_OVERDUE\_SUM | 当前逾期总额,贷款汇总信息 |
| MAX\_RENEGE\_COUNT\_24 | 近24个月最大连续违约次数 |
| SUM\_OVERDUE\_180 | 逾期180天以上未归还贷款本金之和 |
| IS\_OVERDUE | 当前贷款/信用卡是否有逾期 |
| WJQ\_COUNT | 安融征信正常还款账户未结清笔数 |
| JQ\_COUNT | 安融征信正常还款账户已结清笔数 |
| TOTAL\_COUNT | 安融征信正常还款账户小计 |
| EWJQ\_COUNT | 安融征信异常还款账户未结清笔数 |
| EJQ\_COUNT | 安融征信异常还款账户已结清笔数 |
| ETOTAL\_COUNT | 安融征信异常还款账户小计 |
| APPLYING\_COUNT | 安融征信待审笔数 |
| APPLY\_PASSED\_COUNT | 安融征信通过笔数 |
| APPLY\_REJECT\_COUNT | 安融征信拒绝笔数 |
| APPLY\_TOTAL\_COUNT | 安融征信笔数总计 |
| query\_times | 安融征信查询次数 |
| apply\_times | 安融征信申请次数 |

**收入信息相关特征：23个**

|  |  |
| --- | --- |
| 字段名称 | 中文名称 |
| LATEST\_MONTH\_INCOME | 最近一个月工资卡/储蓄卡流水 |
| N1 | N-1个月工资卡/储蓄卡流水 |
| N2 | N-2个月工资卡/储蓄卡流水 |
| N3 | N-3个月工资卡/储蓄卡流水 |
| N4 | N-4个月工资卡/储蓄卡流水 |
| N5 | N-5个月工资卡/储蓄卡流水 |
| MONTH\_INCOME | 月认定收入 |
| END\_BALANCE | 账户期末余额 |
| card\_cnt | 流水银行卡张数 |
| CASH\_FLOW\_TYPE | 流水类型种类个数 |
| CONSUMPTION\_HABITS | 花费习惯种类个数 |
| HAS\_INSURANCE | 是否有社保公积金 |
| HAS\_ACCUMULATION\_FUND | 是否匹配到公积金信息 |
| FUND\_PAYMENT\_PERIOD | 公积金缴费年限 |
| ORGANIZATION\_PORTION | 公积金单位缴存比例 |
| AMOUNT | 公积金月缴存额 |
| LAST\_FUND\_PAYMENT\_MON | 公积金最近缴费期距今月份 |
| FEE\_MONTHS | 社保已缴纳时间 |
| MONTH\_PAYMENT | 社保月缴纳金额 |
| diff | 流水最大差距 |
| mean | 流水均值 |
| maxnump | 流水最大连续断续时间 |
| num | 流水断续次数 |

**评分相关特征：5个**

|  |  |
| --- | --- |
| 字段名称 | 中文名称 |
| CREDIT\_GRADE | FICO评价出的总分 |
| CREDIT\_LEVEL | 信用风险等级 |
| GRADE\_VERSION | 评分版本 |
| CAUTION | 欺诈预警 |
| SCORE | 欺诈分数 |

**申请信息相关特征：9个**

|  |  |
| --- | --- |
| 字段名称 | 中文名称 |
| APPLY\_MAX\_AMOUNT | 最大申请额度 |
| LONG\_REPAYMENT\_TERM | 申请最长还款期限 |
| CHANNEL\_SOURCE | 渠道来源 |
| KNOWN\_CREDITEASE | 了解宜信渠道 |
| MONEY\_SOURCE | 还款来源 |
| ACCEPT\_MOTH\_REPAY | 最高月还款 |
| LOAN\_COUNT | 第几笔贷款 |
| LOAN\_TYPE | 贷款类别 |
| LOAN\_PURPOSE | 借款用途 |

**宜信历史信息：5个**

|  |  |
| --- | --- |
| 字段名称 | 中文名称 |
| MAXOVER\_DAYS | 在宜信最大逾期天数 |
| MAXOVER\_AMOUNT | 在宜信最大逾期金额 |
| MAX\_TOTAL\_OVERDUE\_DAYS | 在宜信最大累计逾期天数 |
| refuse\_times | 被宜信拒贷次数 |
| default\_times | 在宜信违约次数 |

## 4.3 模型特征处理

4.3.1 数据清洗

清理数据中格式’null’或者’..’的数据，将其转换为null

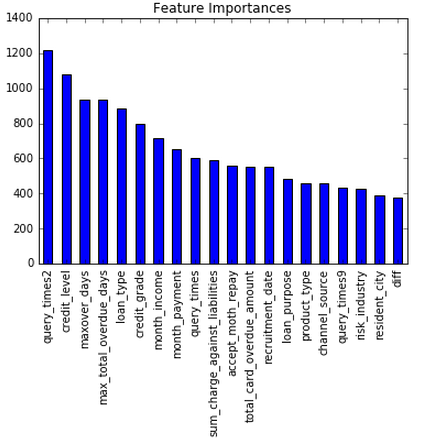
4.3.2 特征筛选

1.根据XGBOOST模型筛选出特征排名靠前的前20特征

即[query\_times2,credit\_level,maxover\_days,max\_total\_overdue\_days,loan\_type,credit\_grade,month\_income,month\_payment,query\_times,sum\_charge\_against\_liabilities,accept\_moth\_repay,total\_card\_overdue\_amount,recruitment\_date,loan\_purpose,product\_type,channel\_source,query\_times9,risk\_industry,resident\_city,diff]

XGBOOST模型参数:

|  |  |
| --- | --- |
| column\_name | value |
| booster | gbtree |
| objective | binary:logistic |
| eval\_metric | auc |
| scale\_pos\_weight | 1.367 |
| max\_depth | 6 |
| subsample | 0.75 |
| colsample\_bytree | 0.75 |
| eta | 0.01 |
| alpha | 0.01 |
| lambda | 1 |



1.计算各个变量间的相关性系数，筛选corr>0.6的相关变量的pair对，对于每个pair对按照feature importances从大大小排序，删除其中较小的一个特征。（图1是各个变量间的相关性系数，图2是相关性系数大于0.6的pair对）

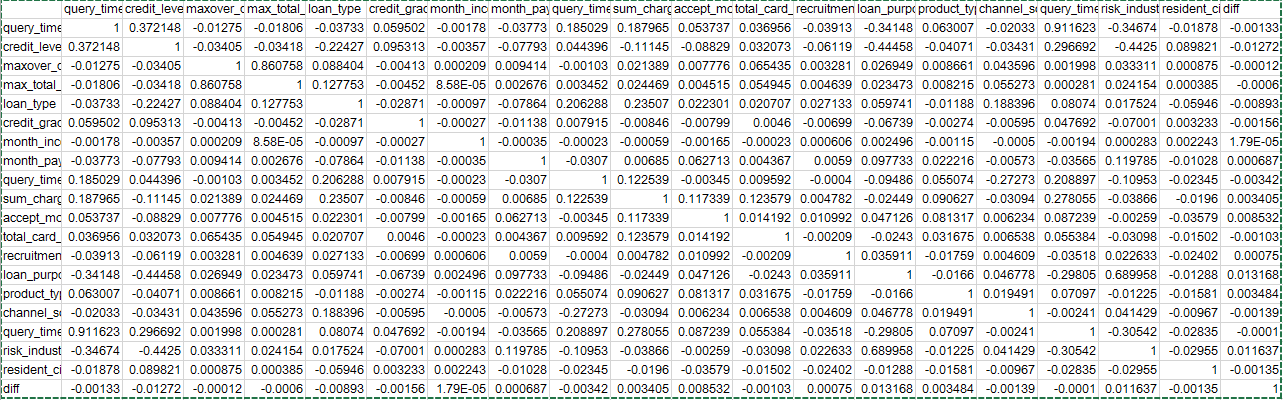


图1 各个变量间的相关性系数

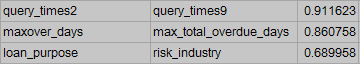


图2 相关性系数大于0.6的变量

综上：筛选后的特征为：

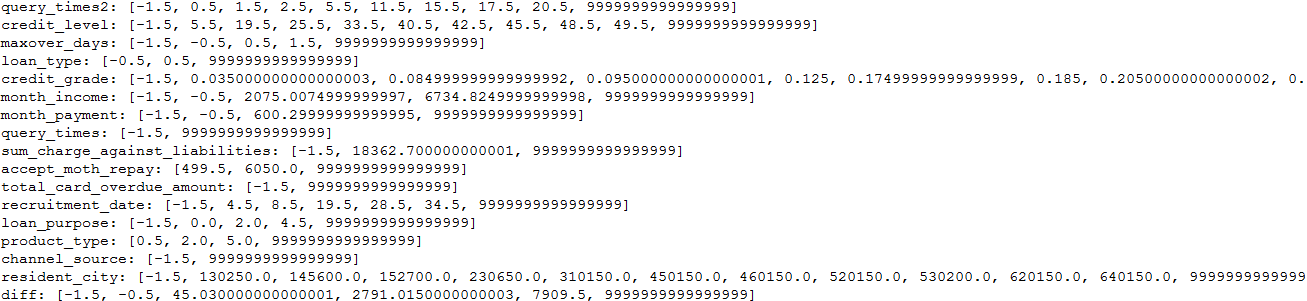
[query\_times2,credit\_level,maxover\_days,loan\_type,credit\_grade,month\_income,month\_payment,query\_times,sum\_charge\_against\_liabilities,accept\_moth\_repay,total\_card\_overdue\_amount,recruitment\_date,loan\_purpose,product\_type,channel\_source, resident\_city,diff]，删除了特征

query\_times9、max\_total\_overdue\_days、risk\_industry

总共剩余17个特征

4.3.2 特征分bin

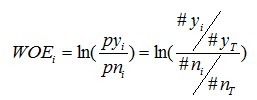
对于连续性变量进行分箱操作



4.3.3 指标变量的woe化

woe定义：woe即对原始自变量的一种编码形式。

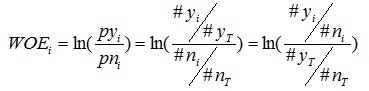
1.对一个变量进行WOE编码，首先需要把这个变量进行分组处理（包括离散化、分等）分组后，对于第i组，WOE的计算公式如下：



其中，pyi是这个组中响应客户（风险模型中，对应的是违约客户，总之，指的是模型中预测变量取值为“是”或者说1的个体）占所有样本中所有响应客户的比例，pni是这个组中未响应客户占样本中所有未响应客户的比例，#yi是这个组中响应客户的数量，#ni是这个组中未响应客户的数量，#yT是样本中所有响应客户的数量，#nT是样本中所有未响应客户的数量。

即WOE表示的实际上是“当前分组中响应客户占所有响应客户的比例”和“当前分组中没有响应的客户占所有没有响应的客户的比例”的差异。

对这个公式做一个简单变换，可以得到：



变换以后我们可以看出，WOE也可以这么理解，他表示的是当前这个组中响应的客户和未响应客户的比值，和所有样本中这个比值的差异。这个差异是用这两个比值的比值，再取对数来表示的。WOE越大，这种差异越大，这个分组里的样本响应的可能性就越大，WOE越小，差异越小，这个分组里的样本响应的可能性就越小。

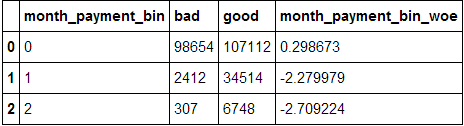


图 3 month\_payment\_bin的woe结果

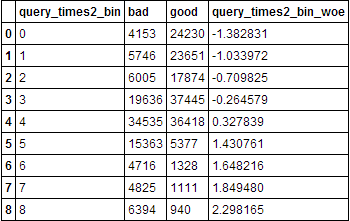


图 4 query\_times2\_bin的woe结果

综上：通过特征选择，处理特征间相关性较大的特征，特征的分bin操作以及特征的woe化，得到最终的特征矩阵。

# 5. 模型实验结果

## 5.1 项目应用的模型

项目采用LR模型对信审数据进行训练和预测,其中训练集共249747条数据，正例数据148374条，负例101373条；测试集共10079条数据，正例数据7019条，负例3060条。（加终审辅助标签）

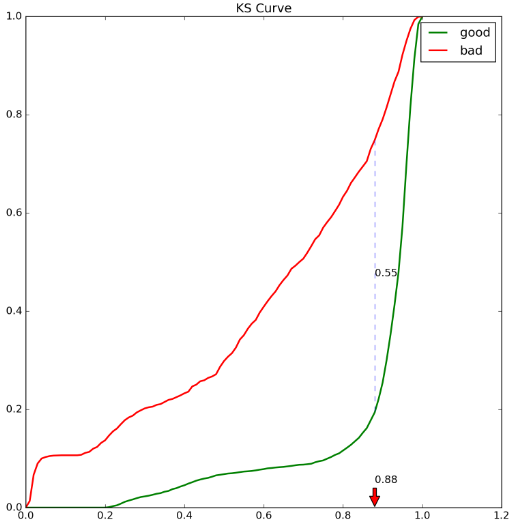
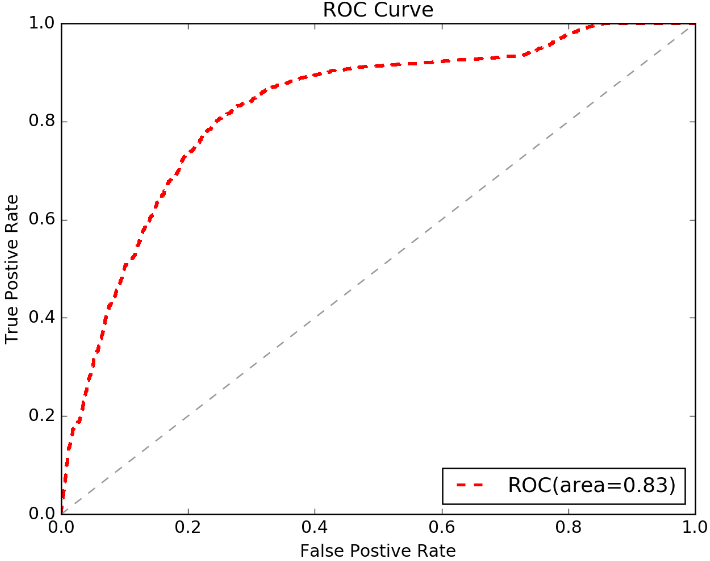
模型结果表：（模型：默认的LR模型）

|  |  |  |
| --- | --- | --- |
| 模型 | 测试集上效果(ks值) | 测试集上效果(auc值) |
| LR | 0.5548 | 82.83 |

LR模型参数:

|  |  |
| --- | --- |
| column\_name | value |
| penalty | l2 |
| C | 10 |
| solver | sag |

**测试集ks和roc曲线如下：**



分析：

ROC曲线下的面积(AUC值)越大，模型对正负样本的区分能力越强；测试集KS曲线；KS曲线中两条累积分布曲线之间的最大Gap(KS值)越大，模型对正负样本的区分能力越强。测试集的AUC为0.83，KS值为0.55，模型达到可用标准。

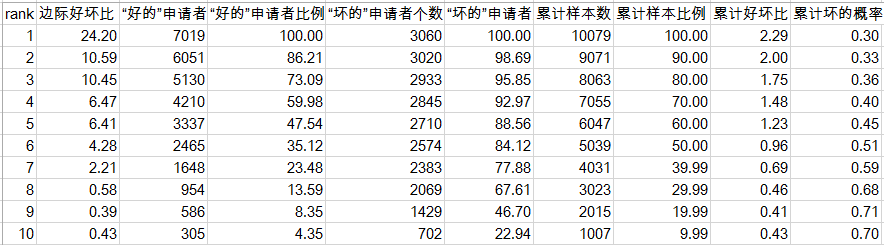
1.申请风险模型的提升指数计算

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 申请者top N | 样本个数 | “好样本”个数 | “好样本”占比 | “好样本”提升指数 |
| Top5% | 503 | 484 | 0.9622 | 138.17% |
| Top 10% | 1007 | 967 | 0.9603 | 137.89% |
| Top 15% | 1511 | 1417 | 0.9378 | 134.66% |

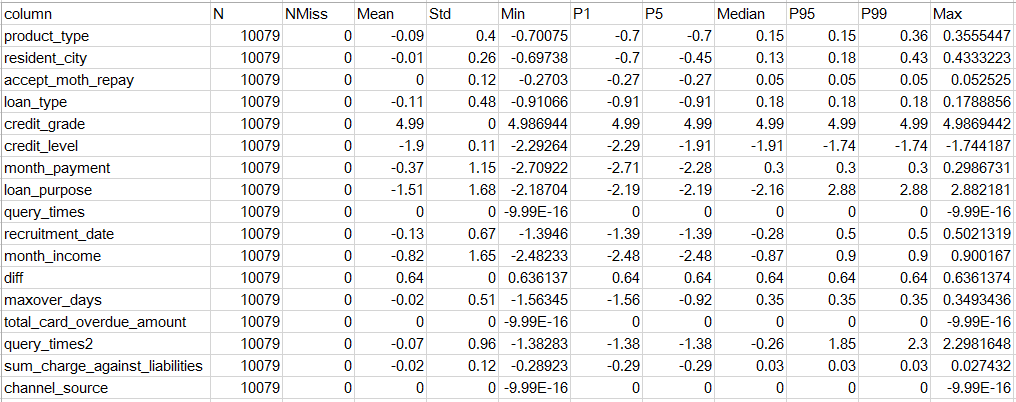
测试集（2017.03.01 - 2017.03.09）：

注：好样本提升指数=（topN好样本占比-总体数据集中好样本占比）/总体数据集好样本占比

2. 申请风险评分分布表

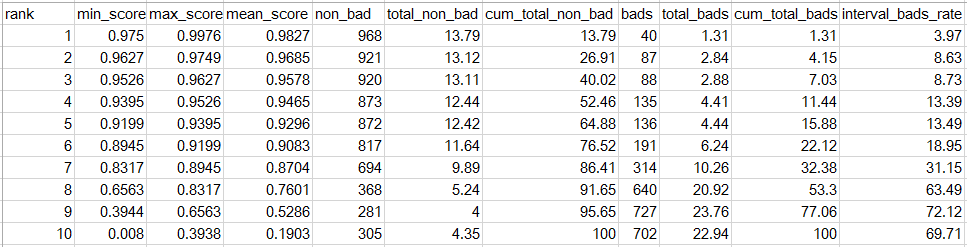


1. 统计测试集的woe后的基本信息



指标说明：N表示样本个数、NMiss表示缺失值个数、Mean 均值、std标准差、Min最小值、P1 1%分位数、P5 5%分位数、Median 中位数、P95 95%分位数、P99 99%分位数、Max 最大值。

4.将测试集按照得分排序，并切分成10份子集，分别统计每份子集的min score,max score,mean score,正样本数，负样本数，以及累计占比等指标



指标说明：rank 分组数、min\_score当前分组的最小得分、max\_score 最大得分、mean\_score平均得分、non\_bad “好的”样本个数、total\_non\_bad “好的”样本占比、cum\_total\_non\_bad “好的”样本累计占比、bads“坏的”样本个数、total\_bads “坏的”样本占比、cum\_total\_bads “坏的”样本累计占比、interval\_bads\_rate 当前分组中“坏的”样本占比。

5.0将测试集按照得分排序，并切分成10份子集，分别统计每份子集以及总体各列特征的均值

