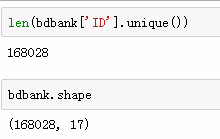
# 一、数据质量分析

## 1.1 数据质量检查

### (1) 主键唯一性检查



主键没有重复情况

### (2) 缺失值与单一值检查

其中标为黄色的特征为单一特征，共6个，后续建模考虑剔除。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | type | 缺失大于50% | 是否单一值 | COUNT |
| bidNormalScoreExp3 | object | 否 | 是 | 160786 |
| creditValidityDays | object | 否 | 是 | 160786 |
| phoneNormalScoreExp3 | object | 否 | 是 | 160786 |
| prcidNormalScoreExp3 | object | 否 | 是 | 160786 |
| risCode | object | 否 | 是 | 160786 |
| zxCreCardStatus2 | object | 否 | 是 | 160786 |
| ID | object | 否 | 否 | 168028 |
| CUST\_NAME | object | 否 | 否 | 168028 |
| ID\_CARD\_NO | object | 否 | 否 | 168028 |
| MOBILE\_NO | int64 | 否 | 否 | 168028 |
| BANK\_CARD\_NO | object | 否 | 否 | 168028 |
| CREATED\_DATE | datetime64[ns] | 否 | 否 | 168028 |
| UPDATED\_DATE | datetime64[ns] | 否 | 否 | 168028 |
| TRANSACTION\_TYPE | object | 否 | 否 | 168028 |
| TRANSACTION\_ID | object | 否 | 否 | 160786 |
| ORDER\_NO | object | 否 | 否 | 168028 |
| INIT\_AMOUNT | float64 | 否 | 否 | 168028 |
| ANNUAL\_INT\_RATE | float64 | 否 | 否 | 160786 |
| CASH\_AMOUNT | float64 | 是 | 否 | 72137 |
| AFTER\_LMT | float64 | 是 | 否 | 7242 |
| BEFORE\_LMT | float64 | 是 | 否 | 7242 |
| PRE\_A\_CUST\_SEG | float64 | 是 | 否 | 0 |
| blackListType | object | 否 | 否 | 160786 |
| cashAmount | object | 是 | 否 | 72137 |
| compreAnnualInterestRate | object | 是 | 否 | 72137 |
| dailyInterestRate | object | 是 | 否 | 72137 |
| dailyPenaltyRate | object | 是 | 否 | 72137 |
| dxmAScore | object | 否 | 否 | 160786 |
| holmesAgentScore | object | 否 | 否 | 160786 |
| holmesBlackScore | object | 否 | 否 | 160786 |
| idpicture0 | object | 是 | 否 | 39261 |
| idpicture1 | object | 是 | 否 | 39261 |
| ifIdPicture | object | 是 | 否 | 39261 |
| ifLiving | object | 是 | 否 | 39261 |
| initialAmount | object | 否 | 否 | 160786 |
| livingPhoto | object | 是 | 否 | 39261 |
| ocrExpdate | object | 是 | 否 | 39261 |
| ocrNation | object | 是 | 否 | 39261 |
| ocrSignOrg | object | 是 | 否 | 39261 |
| orderId | object | 是 | 否 | 72137 |
| photowithidcard | object | 是 | 否 | 39261 |
| preACustSeg | object | 否 | 否 | 160786 |
| r21P12FinallyInterest | object | 否 | 否 | 160786 |
| reasonCode | object | 否 | 否 | 160786 |
| reasonMsg | object | 否 | 否 | 160786 |
| taxMonthlyIncomeSection | object | 否 | 否 | 160786 |
| term | object | 是 | 否 | 72137 |
| unionLoanUsed | object | 是 | 否 | 72137 |
| usedTotalLimit | float64 | 是 | 否 | 72137 |
| zx1MthDelqMaxAmt | object | 否 | 否 | 160786 |
| zxAccountCnt | object | 否 | 否 | 160786 |
| zxAccountNowTotalDelqAmt | object | 否 | 否 | 160786 |
| zxBadBebtsBalance | object | 否 | 否 | 160786 |
| zxBadBebtsCnt | object | 否 | 否 | 160786 |
| zxBlanceShouldPayAmt | object | 否 | 否 | 160786 |
| zxCareer | object | 否 | 否 | 160786 |
| zxCommerHouseLoanBlance | object | 否 | 否 | 160786 |
| zxCreCardAcctNowTotalDelqAmt | object | 否 | 否 | 160786 |
| zxCreCardDelqAccoutCnt | object | 否 | 否 | 160786 |
| zxCreCardDelqMaxAmt | object | 否 | 否 | 160786 |
| zxCreCardMaxMonthsCnt | object | 否 | 否 | 160786 |
| zxCreCardNowShouldPayAmt | object | 否 | 否 | 160786 |
| zxCreCardStatus1 | object | 否 | 否 | 160786 |
| zxCreCardStatus3 | object | 否 | 否 | 160786 |
| zxCreCardStatus4 | object | 否 | 否 | 160786 |
| zxCreCardStatus5 | object | 否 | 否 | 160786 |
| zxCreCardUseCreditRate | object | 否 | 否 | 160786 |
| zxCreditAvgLimit | object | 否 | 否 | 160786 |
| zxCreditBal | object | 否 | 否 | 160786 |
| zxCreditCardAccuDelqCnt | object | 否 | 否 | 160786 |
| zxCreditCardDelqStatus24Months | object | 否 | 否 | 160786 |
| zxCreditCardDelqStatus30Dayyues | object | 否 | 否 | 160786 |
| zxCreditCardDelqStatus30To60Days | object | 否 | 否 | 160786 |
| zxCreditCardDelqStatus60To90Days | object | 否 | 否 | 160786 |
| zxCreditCardDelqStatus6To12Months | object | 否 | 否 | 160786 |
| zxCreditCardDelqStatus90To180Days | object | 否 | 否 | 160786 |
| zxCreditCardRcyDelqDate | object | 否 | 否 | 160786 |
| zxCreditCardRcyUpdateDate | object | 否 | 否 | 160786 |
| zxCreditRetCode | object | 否 | 否 | 160786 |
| zxDelqCnt | object | 否 | 否 | 160786 |
| zxDomicile | object | 否 | 否 | 160786 |
| zxEducation | object | 否 | 否 | 160786 |
| zxFirstCreCardDate | object | 否 | 否 | 160786 |
| zxFirstLoanDate | object | 否 | 否 | 160786 |
| zxFirstLoanMonth | object | 否 | 否 | 160786 |
| zxGuaranteeCnt | object | 否 | 否 | 160786 |
| zxHouseFundCompanyRate | object | 否 | 否 | 160786 |
| zxHouseFundMonthPayAmt | object | 否 | 否 | 160786 |
| zxHouseFundMonthShouldPayAmt | object | 否 | 否 | 160786 |
| zxHouseFundPayLastMonth | object | 否 | 否 | 160786 |
| zxHouseFundPayStatus | object | 否 | 否 | 160786 |
| zxHouseFundRecordUpdate | object | 否 | 否 | 160786 |
| zxHouseFundSelfRate | object | 否 | 否 | 160786 |
| zxHouseLoan1MAvgPayAmt | object | 否 | 否 | 160786 |
| zxIssuBankCnt | object | 否 | 否 | 160786 |
| zxLoanBalCnt | object | 否 | 否 | 160786 |
| zxLoanBalanceAmt | object | 否 | 否 | 160786 |
| zxLoanCompanyBalCnt | object | 否 | 否 | 160786 |
| zxLoanDelqStatus30Dayyues | object | 否 | 否 | 160786 |
| zxLoanDelqStatus30To60Days | object | 否 | 否 | 160786 |
| zxLoanDelqStatus60To90Days | object | 否 | 否 | 160786 |
| zxLoanDelqStatus6To12Months | object | 否 | 否 | 160786 |
| zxLoanDelqStatus90To180Days | object | 否 | 否 | 160786 |
| zxLoanRcyDelqDate | object | 否 | 否 | 160786 |
| zxLoanRcyUpdateDate | object | 否 | 否 | 160786 |
| zxMaritalStatus | object | 否 | 否 | 160786 |
| zxMaxCreditAmt | object | 否 | 否 | 160786 |
| zxMaxMonths | object | 否 | 否 | 160786 |
| zxMaxUpdated | object | 否 | 否 | 160786 |
| zxMinCreditAmt | object | 否 | 否 | 160786 |
| zxPermanentAddress | object | 否 | 否 | 160786 |
| zxPersonalLoanCnt | object | 否 | 否 | 160786 |
| zxRcy12MCreCardMaxDelqPeriod | object | 否 | 否 | 160786 |
| zxRcy1MCreCardQueryCnt | object | 否 | 否 | 160786 |
| zxRcy1MCreCardQueryCompanyCnt | object | 否 | 否 | 160786 |
| zxRcy1MLoanQueryCnt | object | 否 | 否 | 160786 |
| zxRcy1MLoanQueryCompanyCnt | object | 否 | 否 | 160786 |
| zxRcy24MAcctMaxDelqPeriod | object | 否 | 否 | 160786 |
| zxRcy24MCreCardMaxDelqPeriod | object | 否 | 否 | 160786 |
| zxRcy24MCreCardTotalDelqCnt | object | 否 | 否 | 160786 |
| zxRcy3MCreCardQueryCnt | object | 否 | 否 | 160786 |
| zxRcy3MCreCardQueryCompanyCnt | object | 否 | 否 | 160786 |
| zxRcy3MLoanQueryCnt | object | 否 | 否 | 160786 |
| zxRcy3MLoanQueryCompanyCnt | object | 否 | 否 | 160786 |
| zxRcy6MAvgPayAmt | object | 否 | 否 | 160786 |
| zxRcy6MAvgUseCreditAmt | object | 否 | 否 | 160786 |
| zxRcy6MCreCardMaxDelqPeriod | object | 否 | 否 | 160786 |
| zxRcy6MCreCardQueryCnt | object | 否 | 否 | 160786 |
| zxRcy6MCreCardQueryCompanyCnt | object | 否 | 否 | 160786 |
| zxRcy6MLoanQueryCnt | object | 否 | 否 | 160786 |
| zxRcy6MLoanQueryCompanyCnt | object | 否 | 否 | 160786 |
| zxRcyQueryDate | object | 否 | 否 | 160786 |
| zxRestDisposalCnt | object | 否 | 否 | 160786 |
| zxSex | object | 否 | 否 | 160786 |
| zxTotalCreditAmt | object | 否 | 否 | 160786 |
| zxUseCreditAmt | object | 否 | 否 | 160786 |
| zxWarrantLoanCnt | object | 否 | 否 | 160786 |
| zxWarrantLoanRate1 | object | 否 | 否 | 160786 |
| zxWarrantLoanType | object | 否 | 否 | 160786 |
| zxWarrantLoanTypeUnnormal | object | 否 | 否 | 160786 |

(3) 特征清单



## 1.2 数据探索

### (1) 申请数量时间分布

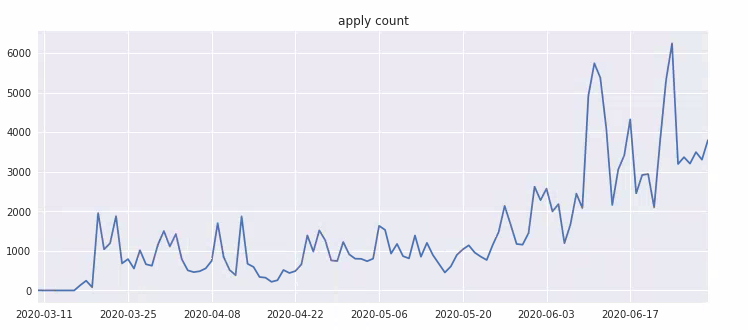


图1-1

从按时间分布的申请数量走势来看，申请客户数量随着时间的推移逐渐增多，在6月份达到顶峰，但我们也可以发现，申请数量在走势呈现出剧烈的震荡形式。

具体每个月的申请数量如表1-3所示

表1-3

|  |  |  |
| --- | --- | --- |
| 月份 | 上旬 | 下旬 |
| 3 | 11 | 13537 |
| 4 | 12697 | 11145 |
| 5 | 15364 | 16949 |
| 6 | 44426 | 53899 |

### (2) 申请数量地区分布

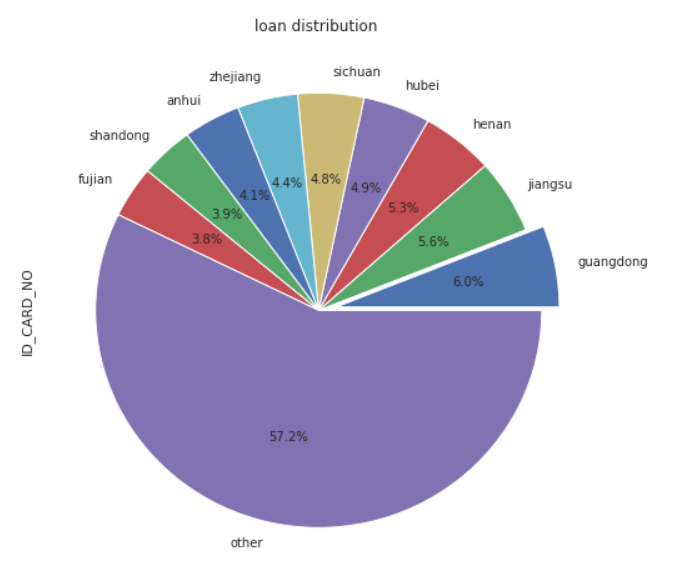


图1-2

申请数量地区分布显示，目前申请客户主要集中子广东、江苏、河南、湖北、四川、浙江、安徽、山东、福建等地，其中申请数量最多的地区为广东省，占比6%

### (3) 申请数量年龄分布

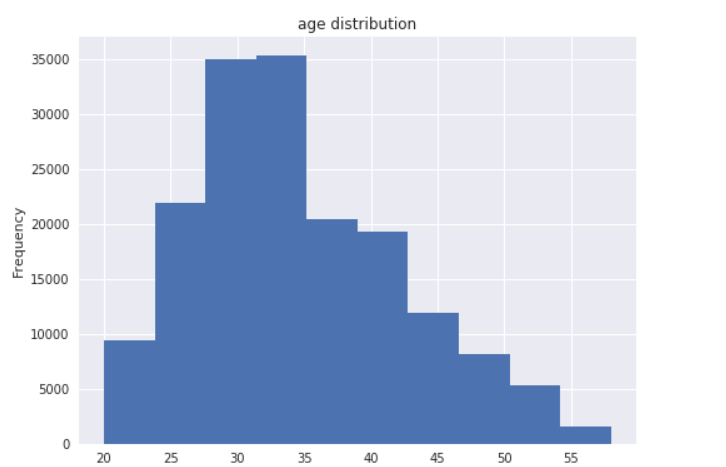
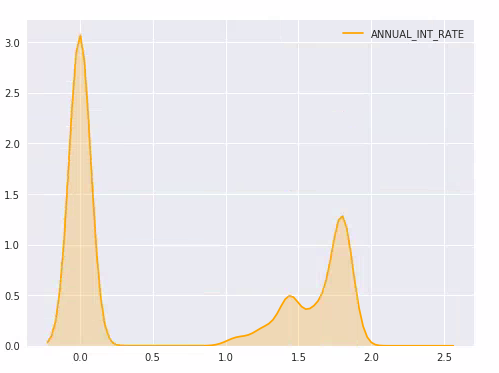


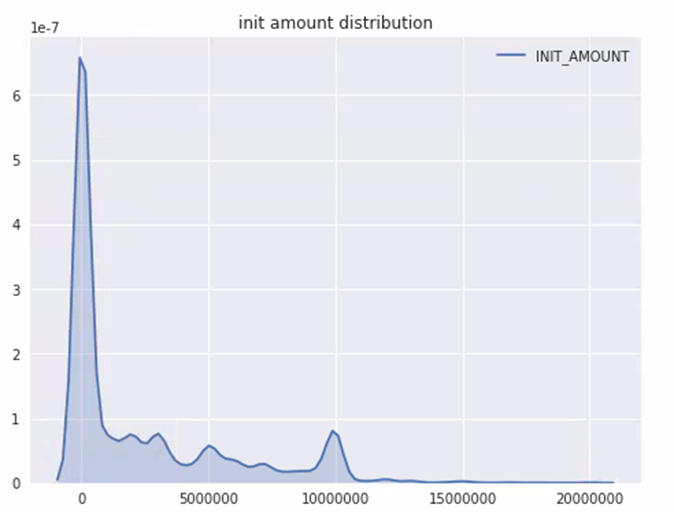
图 1-3

申请数量年龄分布显示申请客户年龄主要集中在20-45区间，具体来看，分布的在30-35区间的申请数量最多。

### (4) 授信年利率分布



### (5) 授信金额分布



## 1.3 逾期滚动情况

根据百度好E贷2020年3月-12月发放的贷款借据，观察2020年11月30日的逾期状态与下个月即2020年12月31日的逾期状态（M0（未逾期）、M1（逾期1-30天）、M2（逾期31-60天）以此类推）的滚动情况。滚动率分析结果如表1-2：

表1-2

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2020年11月-2020年12月逾期滚动率 | | | | | | | | | | |
|  | 12月逾期状态 | | | | | | | | | |
| 11月逾期状态 | M0 | M1 | M2 | M3 | M4 | M5 | M6 | M7 | M8+ | 总计 |
| M0 | 99.80% | 0.19% | 0.01% |  |  |  |  |  |  | 100.00% |
| M1 | 21.30% | 16.65% | 62.05% | 0.00% |  |  |  |  |  | 100.00% |
| M2 | 5.44% | 1.51% | 3.47% | 87.46% | 2.11% |  |  |  |  | 100.00% |
| M3 | 0.69% | 0.00% | 0.00% | 0.23% | 97.24% | 1.84% |  |  |  | 100.00% |
| M4 | 2.07% | 0.00% | 1.04% | 0.52% | 1.04% | 92.23% | 3.11% |  |  | 100.00% |
| M5 | 5.45% | 0.00% | 0.00% | 0.00% | 0.91% | 1.82% | 91.82% | 0.00% |  | 100.00% |
| M6 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 3.57% | 96.43% | 0.00% | 100.00% |
| M7+ | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 100.00% | 100.00% |
| 总计 | 99.52% | 0.22% | 0.11% | 0.07% | 0.05% | 0.02% | 0.01% | 0.00% | 0.00% | 100.00% |

根据滚动率结果表，观察借据在两个月末之间的不同逾期阶段滚动情况如表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2020年11月-2020年12月逾期滚动情况 | | | | |
| 逾期状态 | 更好 | 保持现状 | 更坏 |  |
| M0→ |  | 99.80% | 0.20% | 100.00% |
| M1→ | 21.30% | 16.65% | 62.05% | 100.00% |
| M2→ | 6.95% | 3.47% | 89.58% | 100.00% |
| M3→ | 0.69% | 0.23% | 99.08% | 100.00% |
| M4→ | 3.63% | 1.04% | 95.34% | 100.00% |
| M5→ | 6.36% | 1.82% | 91.82% | 100.00% |
| M6→ | 0.00% | 3.57% | 96.43% | 100.00% |
| M7+→ | 0.00% | 0.00% | 100.00% | 100.00% |

由滚动情况可以看出，逾期状态M1的客户在未来有62.05%的概率继续恶化，21.3%的概率状态变优；逾期状态M2的客户在未来有89.58%的概率继续恶化，有6.95%的概率状态变优; 因此，M2是客户落入不良的关键节点。逾期状态M4虽然有些波动，但总体逾期状态依然在M2后保持较高恶化概率。

基于以上分析，我们的好坏标签定义如下：

|  |  |
| --- | --- |
| 标签 | 定义规则 |
| 好客户 | 逾期天数=0天 |
| 坏客户 | 逾期天数>30天 |
| 灰度客户 | 0<逾期天数≤30天 |

## 1.4 窗口定义

百度好E贷模型表现窗口与观察窗口选择依据账龄分析、贷款期限和逾期率变化趋势确定。截至2020年12月31日，好E贷M2+贷款金额账龄分析表如下：

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| M2+账龄分析 | | | | | | | | | | |
| lon\_mon | **MOB0** | **MOB1** | **MOB2** | **MOB3** | **MOB4** | **MOB5** | **MOB6** | **MOB7** | **MOB8** | **MOB9** |
| 2020-03 | 0.00% | 0.00% | 0.03% | 0.31% | 0.31% | 0.31% | 0.41% | 0.58% | 0.65% | 0.61% |
| 2020-04 | 0.00% | 0.00% | 0.06% | 0.14% | 0.33% | 0.45% | 0.65% | 0.62% | 0.77% | 0.79% |
| 2020-05 | 0.00% | 0.00% | 0.15% | 0.40% | 0.53% | 0.75% | 0.96% | 1.17% | 1.07% |  |
| 2020-06 | 0.00% | 0.00% | 0.15% | 0.37% | 0.66% | 0.96% | 1.22% | 1.20% |  |  |
| 2020-07 | 0.00% | 0.00% | 0.10% | 0.32% | 0.52% | 0.66% | 0.59% |  |  |  |
| 2020-08 | 0.00% | 0.00% | 0.14% | 0.36% | 0.58% | 0.59% |  |  |  |  |
| 2020-09 | 0.00% | 0.00% | 0.09% | 0.28% | 0.20% |  |  |  |  |  |
| 2020-10 | 0.00% | 0.00% | 0.10% | 0.14% |  |  |  |  |  |  |
| 2020-11 | 0.00% | 0.00% | 0.00% |  |  |  |  |  |  |  |
| 2020-12 | 0.00% | 0.00% |  |  |  |  |  |  |  |  |
| 平均M2+逾期率 | 0.00% | 0.00% | 0.09% | 0.29% | 0.45% | 0.62% | 0.77% | 0.89% | 0.83% | 0.70% |

截至2020年12月31日，好e贷每月逾期情况如下：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 累计客户数 | 累计逾期客户数 | 累计客户逾期率 |
| 2020-03-31 | 2094 | 0.0 | 0.00% |
| 2020-04-30 | 7084 | 1.0 | 0.01% |
| 2020-05-31 | 14351 | 15.0 | 0.10% |
| 2020-06-30 | 37744 | 40.0 | 0.11% |
| 2020-07-31 | 68685 | 141.0 | 0.21% |
| 2020-08-31 | 88201 | 309.0 | 0.35% |
| 2020-09-30 | 126381 | 591.0 | 0.47% |
| 2020-10-31 | 197876 | 976.0 | 0.49% |
| 2020-11-30 | 247323 | 1524.0 | 0.62% |
| 2020-12-31 | 283600 | 2249.0 | 0.79% |

M2+贷款金额账龄分析表来看，各个月份发放的贷款随着账龄的增长，风险仍在持续暴露当中，在MOB5时平均逾期率为0.62%。在MOB6逐渐稳定在7%左右。

在观察期选择上，因为百度渠道上线较晚，因此为保证能有较多的坏样本，综合其他渠道，选择2020年3月至2020年6月作为观察期，用此区间的贷款数据构建训练集。同时，在表现期选择上，我们从M2+账龄分析可以看出，在MOB6逾期情况已经有充分的暴露，因此，选择表现期6个月（360天）作为表现窗口

综上，此次建模选取2020年3月-6月作为观察窗口，用于搜集模型开发所需的样本。选取表现窗口为360天，用于确定在观察期间内贷款的客户的好坏表现。