AKULAKU：idpdlold86建模报告

|  |  |
| --- | --- |
| 制作部门 | AKULAKU风控建模 |
| 使用模型 | xgb |
| 模型版本 | v1.0 |
| 制作单位 | AKULAKU |
| 制作时间 | 2020-10-13 |

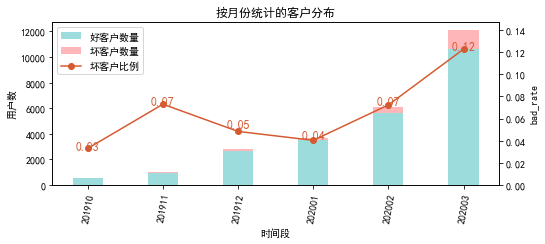
# 一、建模样本

表1 样本概况表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 合作单位 | 样本总量 | 总体坏客户比例 | 训练集测试集起始日期 | 训练集测试集结束日期 | 时间外样本起始日期 | 时间外样本结束日期 |
|  | 26230 | 0.0877 | 201910 | 202002 | 202003 | 202003 |

表2 按月份统计分布表

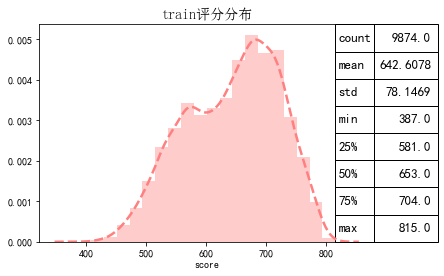
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 201910 | 527 | 18 | 545 | 3.3% | 383.0 | 11.0 | 394.0 | 2.8% | 144.0 | 7.0 | 151.0 | 4.6% |
| 201911 | 928 | 73 | 1001 | 7.3% | 664.0 | 51.0 | 715.0 | 7.1% | 264.0 | 22.0 | 286.0 | 7.7% |
| 201912 | 2656 | 135 | 2791 | 4.8% | 1877.0 | 92.0 | 1969.0 | 4.7% | 779.0 | 43.0 | 822.0 | 5.2% |
| 202001 | 3520 | 148 | 3668 | 4.0% | 2449.0 | 105.0 | 2554.0 | 4.1% | 1071.0 | 43.0 | 1114.0 | 3.9% |
| 202002 | 5661 | 441 | 6102 | 7.2% | 3931.0 | 311.0 | 4242.0 | 7.3% | 1730.0 | 130.0 | 1860.0 | 7.0% |
| 202003 | 10637 | 1486 | 12123 | 12.3% | 0.0 | 0.0 | 0.0 | 0.0% | 0.0 | 0.0 | 0.0 | 0.0% |

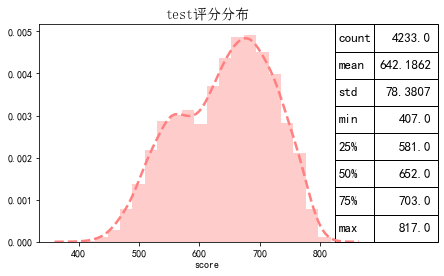


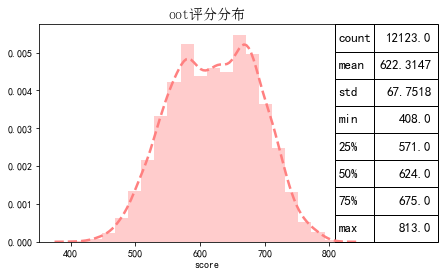
# 二、效果概况

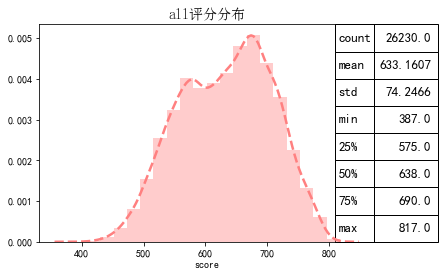
表3 模型效果概况表

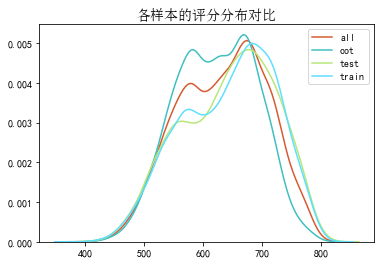
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | train | test | oot | ALL |
| 样本量 | 9874 | 4233 | 12123 | 26230 |
| 坏客户数量 | 570 | 245 | 1486 | 2301 |
| 坏客户比例 | 5.8% | 5.8% | 12.3% | 8.8% |
| KS | 0.524 | 0.498 | 0.239 | 0.356 |
| AUC | 0.839 | 0.809 | 0.668 | 0.742 |





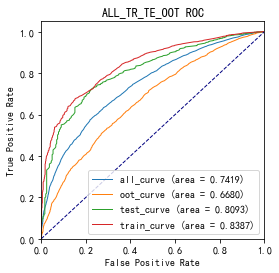
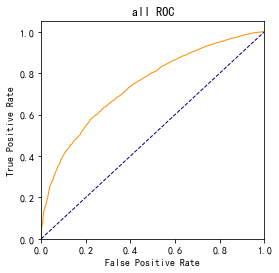
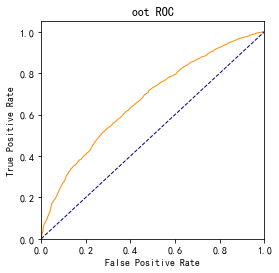
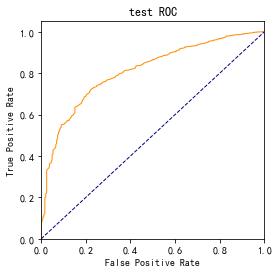
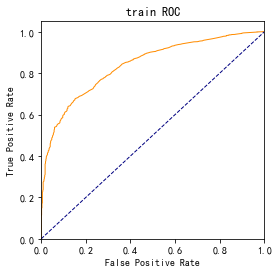






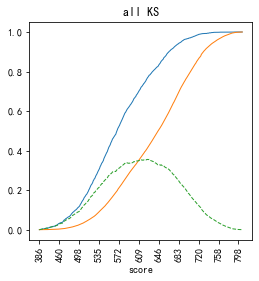
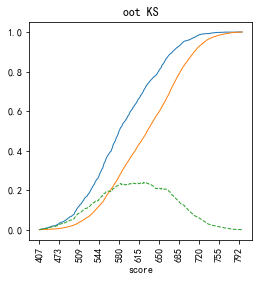
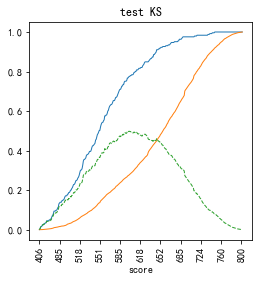
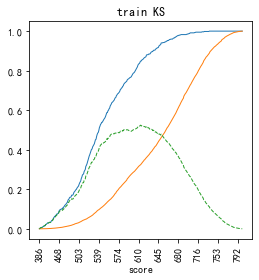
# 三、ROC曲线

ROC（Eceiver operating characteristic curve）：以FPR为横轴，TPR为纵轴，在不同阈值下计算FPR和TPR的值画出的图形。ROC曲线同对角线形成的面积越大说明模型排序能力越好。



# 四、KS曲线

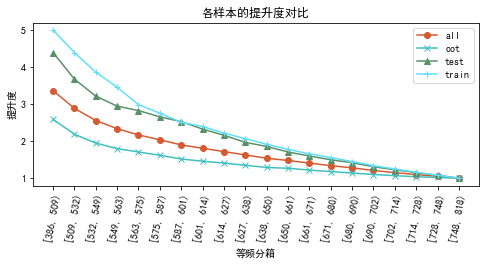
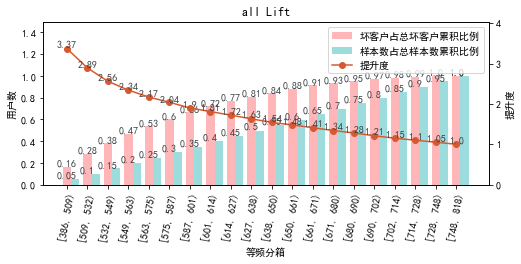
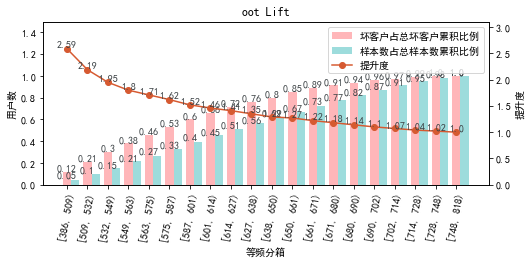
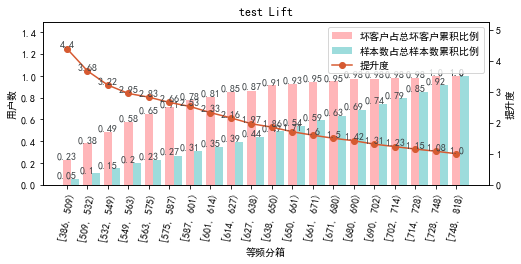
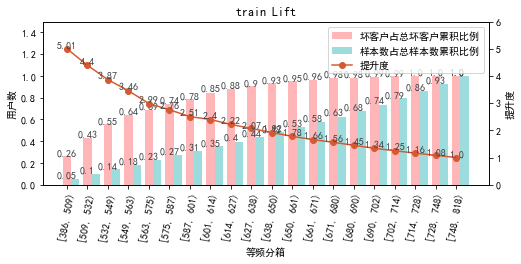
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

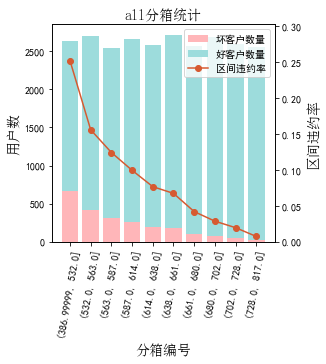
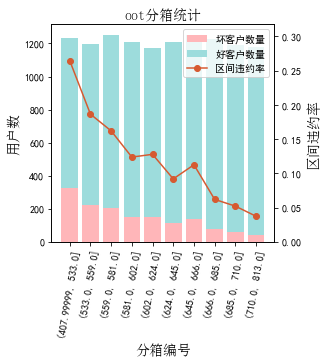
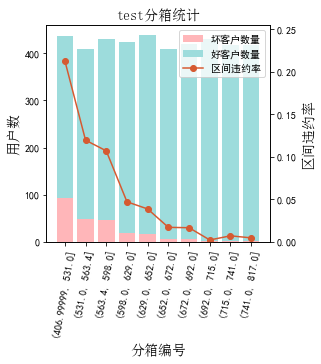
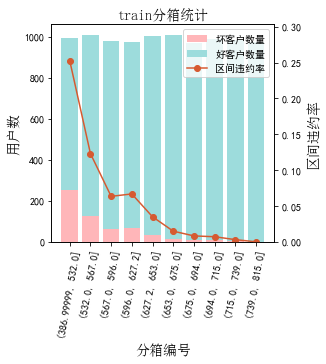
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

将评分等频分为20份，坏客户占总坏客户的累积比例，比样本数占总体样本数的累积比例，得提升度Lift，即该模型抓取坏客户的能力是随机选择的多少倍。



# 六、等频分箱

将训练集、测试集、总体分数均分为数量相同的10个区间，每个区间对应的好客户、坏客户数量及区间违约率。



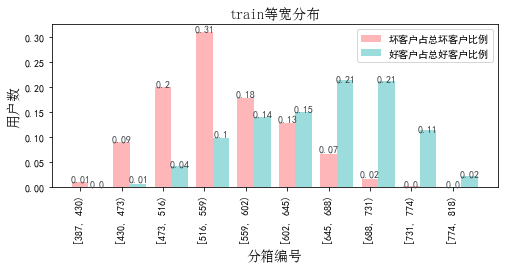
# 七、分数分布

将训练集、测试集分数分为分数间距相同的10个区间，每个区间对应的好客户、坏客户占比及区间违约率。

## 1.train

表4 等宽分箱数据分布表

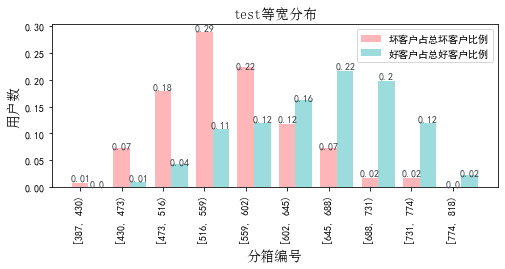
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [387, 430) | 9 | 6 | 3 | 1.1% | 0.0% | 1.1% | 0.0% | 66.7% | 9 | 6 | 0.1% | 66.7% |
| [430, 473) | 113 | 51 | 62 | 8.9% | 0.7% | 10.0% | 0.7% | 45.1% | 122 | 57 | 1.2% | 46.7% |
| [473, 516) | 497 | 114 | 383 | 20.0% | 4.1% | 30.0% | 4.8% | 22.9% | 619 | 171 | 6.3% | 27.6% |
| [516, 559) | 1088 | 177 | 911 | 31.1% | 9.8% | 61.1% | 14.6% | 16.3% | 1707 | 348 | 17.3% | 20.4% |
| [559, 602) | 1405 | 101 | 1304 | 17.7% | 14.0% | 78.8% | 28.6% | 7.2% | 3112 | 449 | 31.5% | 14.4% |
| [602, 645) | 1474 | 73 | 1401 | 12.8% | 15.1% | 91.6% | 43.7% | 5.0% | 4586 | 522 | 46.4% | 11.4% |
| [645, 688) | 2033 | 38 | 1995 | 6.7% | 21.4% | 98.2% | 65.1% | 1.9% | 6619 | 560 | 67.0% | 8.5% |
| [688, 731) | 1974 | 9 | 1965 | 1.6% | 21.1% | 99.8% | 86.2% | 0.5% | 8593 | 569 | 87.0% | 6.6% |
| [731, 774) | 1069 | 1 | 1068 | 0.2% | 11.5% | 100.0% | 97.7% | 0.1% | 9662 | 570 | 97.9% | 5.9% |
| [774, 818) | 212 | 0 | 212 | 0.0% | 2.3% | 100.0% | 100.0% | 0.0% | 9874 | 570 | 100.0% | 5.8% |



## 2.test

表5 等宽分箱数据分布表

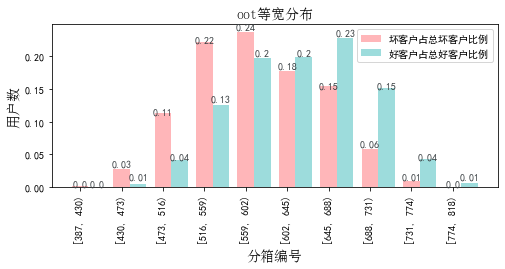
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [387, 430) | 2 | 2 | 0 | 0.8% | 0.0% | 0.8% | 0.0% | 100.0% | 2 | 2 | 0.0% | 100.0% |
| [430, 473) | 52 | 18 | 34 | 7.3% | 0.9% | 8.2% | 0.9% | 34.6% | 54 | 20 | 1.3% | 37.0% |
| [473, 516) | 217 | 44 | 173 | 18.0% | 4.3% | 26.1% | 5.2% | 20.3% | 271 | 64 | 6.4% | 23.6% |
| [516, 559) | 501 | 71 | 430 | 29.0% | 10.8% | 55.1% | 16.0% | 14.2% | 772 | 135 | 18.2% | 17.5% |
| [559, 602) | 532 | 55 | 477 | 22.4% | 12.0% | 77.6% | 27.9% | 10.3% | 1304 | 190 | 30.8% | 14.6% |
| [602, 645) | 680 | 29 | 651 | 11.8% | 16.3% | 89.4% | 44.3% | 4.3% | 1984 | 219 | 46.9% | 11.0% |
| [645, 688) | 885 | 18 | 867 | 7.3% | 21.7% | 96.7% | 66.0% | 2.0% | 2869 | 237 | 67.8% | 8.3% |
| [688, 731) | 794 | 4 | 790 | 1.6% | 19.8% | 98.4% | 85.8% | 0.5% | 3663 | 241 | 86.5% | 6.6% |
| [731, 774) | 483 | 4 | 479 | 1.6% | 12.0% | 100.0% | 97.8% | 0.8% | 4146 | 245 | 97.9% | 5.9% |
| [774, 818) | 87 | 0 | 87 | 0.0% | 2.2% | 100.0% | 100.0% | 0.0% | 4233 | 245 | 100.0% | 5.8% |



## 3.oot

表6 等宽分箱数据分布表

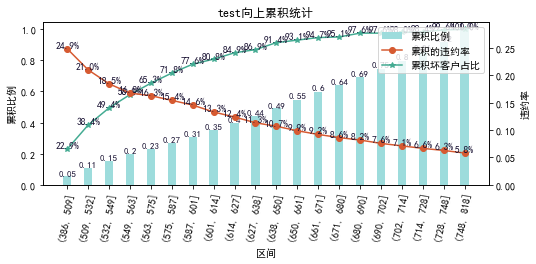
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [387, 430) | 5 | 2 | 3 | 0.1% | 0.0% | 0.1% | 0.0% | 40.0% | 5 | 2 | 0.0% | 40.0% |
| [430, 473) | 96 | 41 | 55 | 2.8% | 0.5% | 2.9% | 0.5% | 42.7% | 101 | 43 | 0.8% | 42.6% |
| [473, 516) | 605 | 169 | 436 | 11.4% | 4.1% | 14.3% | 4.6% | 27.9% | 706 | 212 | 5.8% | 30.0% |
| [516, 559) | 1671 | 329 | 1342 | 22.1% | 12.6% | 36.4% | 17.3% | 19.7% | 2377 | 541 | 19.6% | 22.8% |
| [559, 602) | 2458 | 353 | 2105 | 23.8% | 19.8% | 60.2% | 37.0% | 14.4% | 4835 | 894 | 39.9% | 18.5% |
| [602, 645) | 2387 | 263 | 2124 | 17.7% | 20.0% | 77.9% | 57.0% | 11.0% | 7222 | 1157 | 59.6% | 16.0% |
| [645, 688) | 2654 | 229 | 2425 | 15.4% | 22.8% | 93.3% | 79.8% | 8.6% | 9876 | 1386 | 81.5% | 14.0% |
| [688, 731) | 1704 | 86 | 1618 | 5.8% | 15.2% | 99.1% | 95.0% | 5.0% | 11580 | 1472 | 95.5% | 12.7% |
| [731, 774) | 470 | 13 | 457 | 0.9% | 4.3% | 99.9% | 99.3% | 2.8% | 12050 | 1485 | 99.4% | 12.3% |
| [774, 818) | 73 | 1 | 72 | 0.1% | 0.7% | 100.0% | 100.0% | 1.4% | 12123 | 1486 | 100.0% | 12.3% |

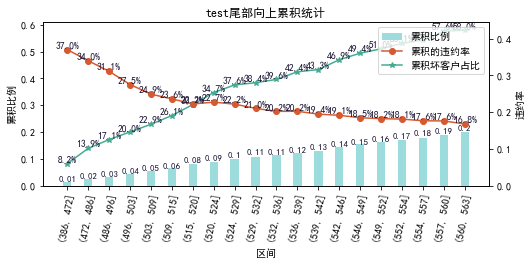


# 八、累积分数分布

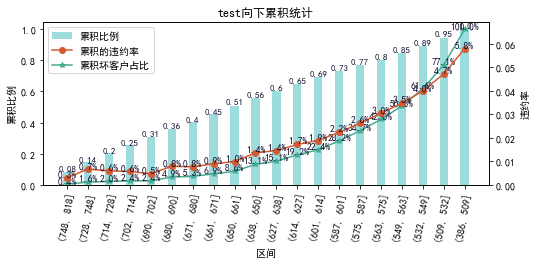
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

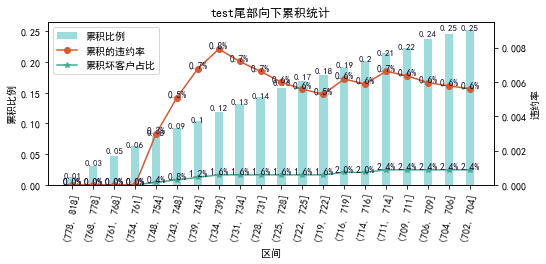
## 1. 向上累积分布





## 2. 向下累积分布





# 九、PSI

PSI（Population stability index）：可衡量测试样本及模型开发样本评分的的分布差异，检验变量的稳定性。psi值越小，变量分布差异越小，越稳定。

表6 PSI计算表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 评分 | train样本量占比 | test样本量占比 | oot\_pct | train\_test\_psi | train\_oot\_psi |
| [386.57, 430.0) | 0.1% | 0.0% | 0.0% | 0.0 | 0.0 |
| [430.0, 473.0) | 1.1% | 1.2% | 0.8% | 0.0 | 0.001 |
| [473.0, 516.0) | 5.0% | 5.1% | 5.0% | 0.0 | 0.0 |
| [516.0, 559.0) | 11.0% | 11.8% | 13.8% | 0.001 | 0.006 |
| [559.0, 602.0) | 14.2% | 12.6% | 20.3% | 0.002 | 0.021 |
| [602.0, 645.0) | 14.9% | 16.1% | 19.7% | 0.001 | 0.013 |
| [645.0, 688.0) | 20.6% | 20.9% | 21.9% | 0.0 | 0.001 |
| [688.0, 731.0) | 20.0% | 18.8% | 14.1% | 0.001 | 0.021 |
| [731.0, 774.0) | 10.8% | 11.4% | 3.9% | 0.0 | 0.071 |
| [774.0, 817.0) | 2.1% | 2.0% | 0.6% | 0.0 | 0.02 |
| sum | 100.0% | 100.0% | 100.0% | 0.005 | 0.155 |

# 十、策略建议