AKULAKU：idpdlold85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 |

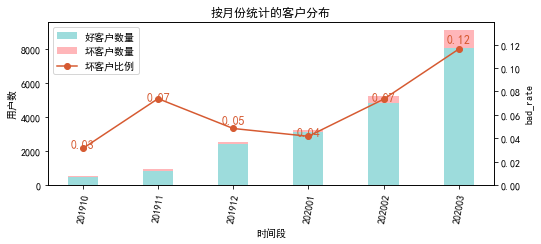
# 一、建模样本

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 |
|  | 21656 | 0.0827 | 201910 | 202002 | 202003 | 202003 |

表2 按月份统计分布表

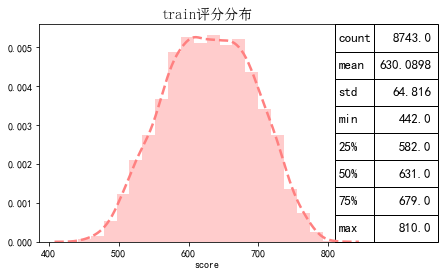
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 201910 | 495 | 16 | 511 | 3.1% | 353.0 | 11.0 | 364.0 | 3.0% | 142.0 | 5.0 | 147.0 | 3.4% |
| 201911 | 853 | 68 | 921 | 7.4% | 605.0 | 48.0 | 653.0 | 7.4% | 248.0 | 20.0 | 268.0 | 7.5% |
| 201912 | 2442 | 124 | 2566 | 4.8% | 1706.0 | 75.0 | 1781.0 | 4.2% | 736.0 | 49.0 | 785.0 | 6.2% |
| 202001 | 3105 | 135 | 3240 | 4.2% | 2155.0 | 94.0 | 2249.0 | 4.2% | 950.0 | 41.0 | 991.0 | 4.1% |
| 202002 | 4868 | 385 | 5253 | 7.3% | 3414.0 | 282.0 | 3696.0 | 7.6% | 1454.0 | 103.0 | 1557.0 | 6.6% |
| 202003 | 8101 | 1064 | 9165 | 11.6% | 0.0 | 0.0 | 0.0 | 0.0% | 0.0 | 0.0 | 0.0 | 0.0% |

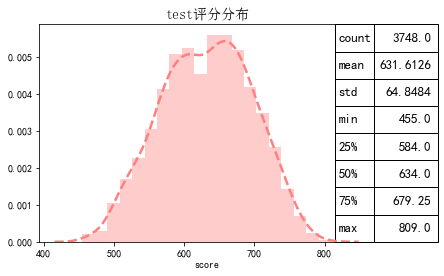


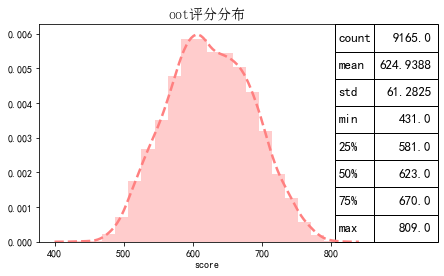
# 二、效果概况

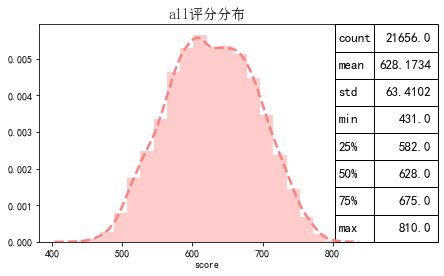
表3 模型效果概况表

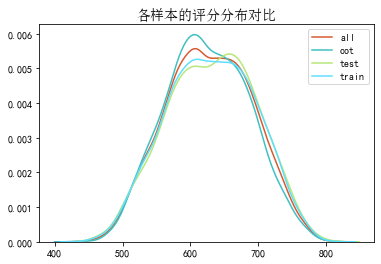
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | train | test | oot | ALL |
| 样本量 | 8743 | 3748 | 9165 | 21656 |
| 坏客户数量 | 510 | 218 | 1064 | 1792 |
| 坏客户比例 | 5.8% | 5.8% | 11.6% | 8.3% |
| KS | 0.442 | 0.4 | 0.282 | 0.34 |
| AUC | 0.786 | 0.765 | 0.689 | 0.729 |





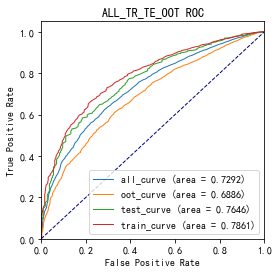
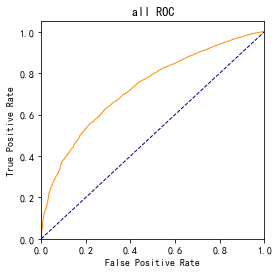
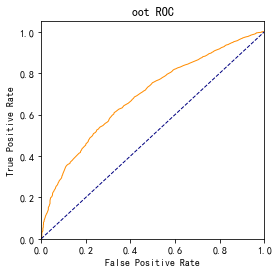
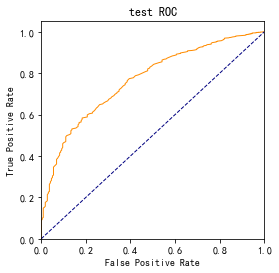
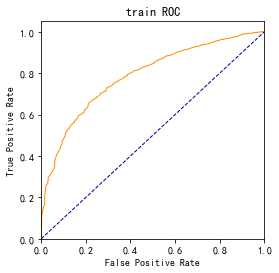






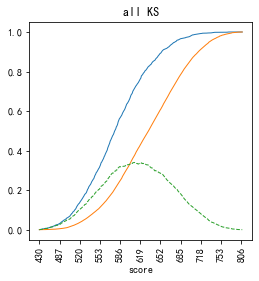
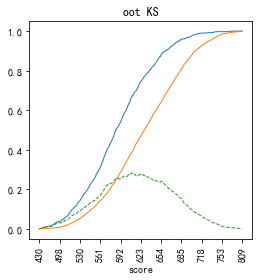
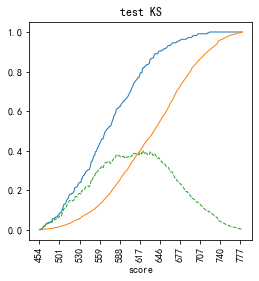
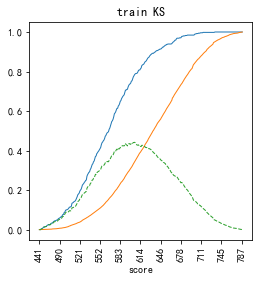
# 三、ROC曲线

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

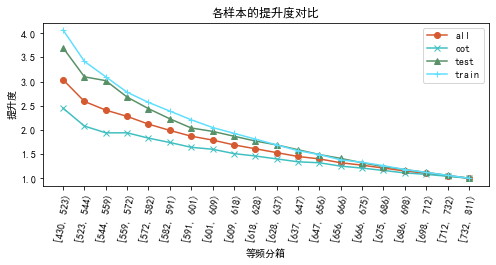
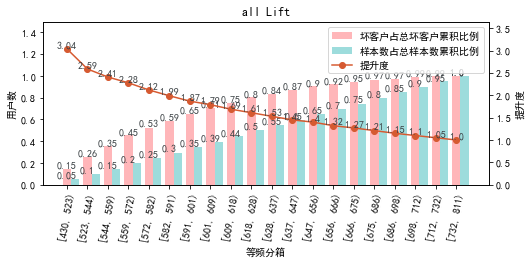
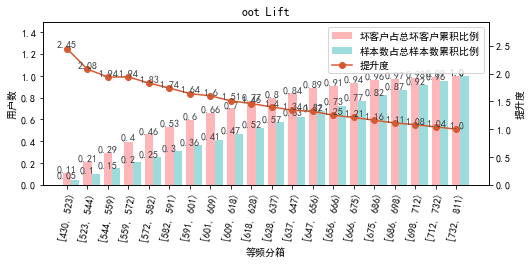
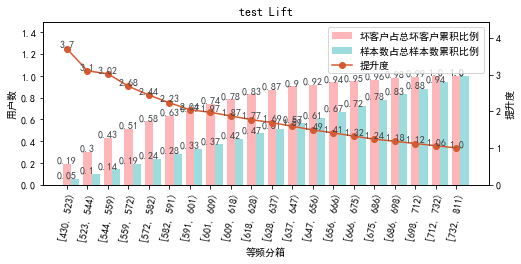
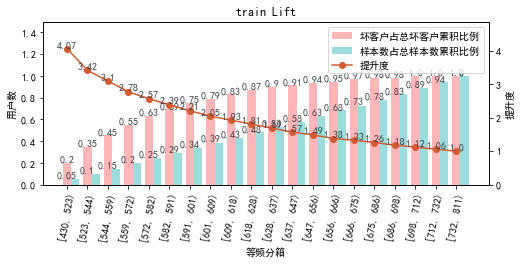
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

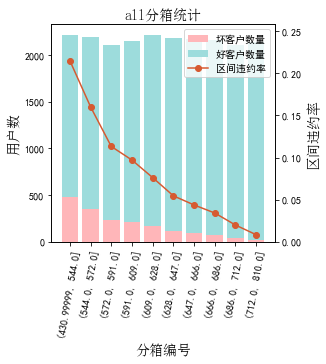
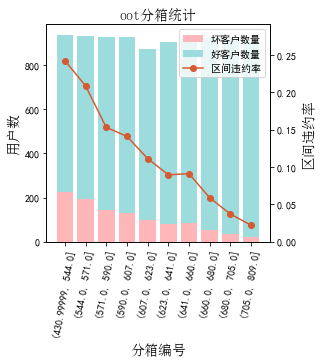
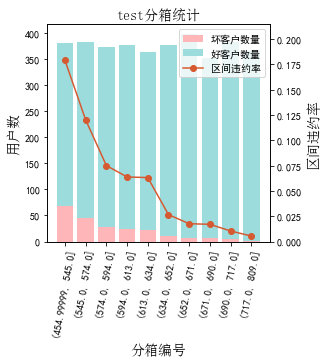
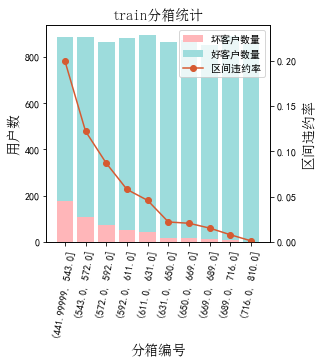
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



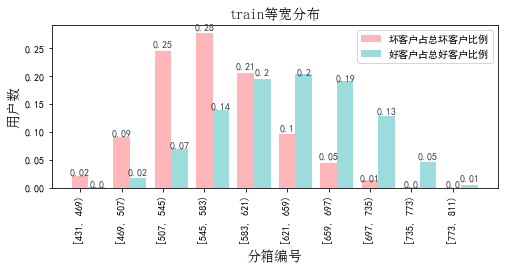
# 七、分数分布

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

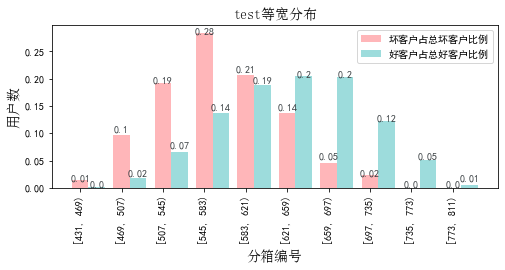
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [431, 469) | 21 | 11 | 10 | 2.2% | 0.1% | 2.2% | 0.1% | 52.4% | 21 | 11 | 0.2% | 52.4% |
| [469, 507) | 191 | 47 | 144 | 9.2% | 1.7% | 11.4% | 1.9% | 24.6% | 212 | 58 | 2.4% | 27.4% |
| [507, 545) | 696 | 125 | 571 | 24.5% | 6.9% | 35.9% | 8.8% | 18.0% | 908 | 183 | 10.4% | 20.2% |
| [545, 583) | 1289 | 142 | 1147 | 27.8% | 13.9% | 63.7% | 22.7% | 11.0% | 2197 | 325 | 25.1% | 14.8% |
| [583, 621) | 1715 | 105 | 1610 | 20.6% | 19.6% | 84.3% | 42.3% | 6.1% | 3912 | 430 | 44.7% | 11.0% |
| [621, 659) | 1729 | 49 | 1680 | 9.6% | 20.4% | 93.9% | 62.7% | 2.8% | 5641 | 479 | 64.5% | 8.5% |
| [659, 697) | 1606 | 23 | 1583 | 4.5% | 19.2% | 98.4% | 81.9% | 1.4% | 7247 | 502 | 82.9% | 6.9% |
| [697, 735) | 1071 | 7 | 1064 | 1.4% | 12.9% | 99.8% | 94.8% | 0.7% | 8318 | 509 | 95.1% | 6.1% |
| [735, 773) | 379 | 1 | 378 | 0.2% | 4.6% | 100.0% | 99.4% | 0.3% | 8697 | 510 | 99.5% | 5.9% |
| [773, 811) | 46 | 0 | 46 | 0.0% | 0.6% | 100.0% | 100.0% | 0.0% | 8743 | 510 | 100.0% | 5.8% |



## 2.test

表5 等宽分箱数据分布表

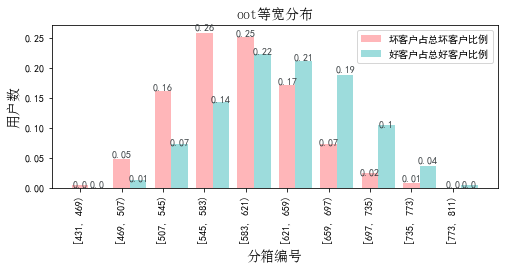
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [431, 469) | 9 | 3 | 6 | 1.4% | 0.2% | 1.4% | 0.2% | 33.3% | 9 | 3 | 0.2% | 33.3% |
| [469, 507) | 83 | 21 | 62 | 9.6% | 1.8% | 11.0% | 1.9% | 25.3% | 92 | 24 | 2.5% | 26.1% |
| [507, 545) | 277 | 42 | 235 | 19.3% | 6.7% | 30.3% | 8.6% | 15.2% | 369 | 66 | 9.8% | 17.9% |
| [545, 583) | 544 | 62 | 482 | 28.4% | 13.7% | 58.7% | 22.2% | 11.4% | 913 | 128 | 24.4% | 14.0% |
| [583, 621) | 713 | 45 | 668 | 20.6% | 18.9% | 79.4% | 41.2% | 6.3% | 1626 | 173 | 43.4% | 10.6% |
| [621, 659) | 752 | 30 | 722 | 13.8% | 20.5% | 93.1% | 61.6% | 4.0% | 2378 | 203 | 63.4% | 8.5% |
| [659, 697) | 730 | 10 | 720 | 4.6% | 20.4% | 97.7% | 82.0% | 1.4% | 3108 | 213 | 82.9% | 6.9% |
| [697, 735) | 440 | 5 | 435 | 2.3% | 12.3% | 100.0% | 94.3% | 1.1% | 3548 | 218 | 94.7% | 6.1% |
| [735, 773) | 179 | 0 | 179 | 0.0% | 5.1% | 100.0% | 99.4% | 0.0% | 3727 | 218 | 99.4% | 5.8% |
| [773, 811) | 21 | 0 | 21 | 0.0% | 0.6% | 100.0% | 100.0% | 0.0% | 3748 | 218 | 100.0% | 5.8% |



## 3.oot

表6 等宽分箱数据分布表

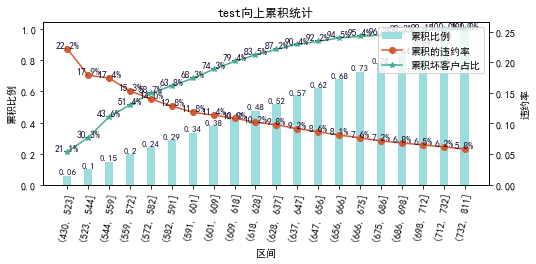
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [431, 469) | 7 | 5 | 2 | 0.5% | 0.0% | 0.5% | 0.0% | 71.4% | 7 | 5 | 0.1% | 71.4% |
| [469, 507) | 163 | 51 | 112 | 4.8% | 1.4% | 5.3% | 1.4% | 31.3% | 170 | 56 | 1.9% | 32.9% |
| [507, 545) | 767 | 171 | 596 | 16.1% | 7.4% | 21.3% | 8.8% | 22.3% | 937 | 227 | 10.2% | 24.2% |
| [545, 583) | 1439 | 275 | 1164 | 25.8% | 14.4% | 47.2% | 23.1% | 19.1% | 2376 | 502 | 25.9% | 21.1% |
| [583, 621) | 2067 | 268 | 1799 | 25.2% | 22.2% | 72.4% | 45.3% | 13.0% | 4443 | 770 | 48.5% | 17.3% |
| [621, 659) | 1897 | 182 | 1715 | 17.1% | 21.2% | 89.5% | 66.5% | 9.6% | 6340 | 952 | 69.2% | 15.0% |
| [659, 697) | 1603 | 78 | 1525 | 7.3% | 18.8% | 96.8% | 85.3% | 4.9% | 7943 | 1030 | 86.7% | 13.0% |
| [697, 735) | 876 | 26 | 850 | 2.4% | 10.5% | 99.2% | 95.8% | 3.0% | 8819 | 1056 | 96.2% | 12.0% |
| [735, 773) | 309 | 8 | 301 | 0.8% | 3.7% | 100.0% | 99.5% | 2.6% | 9128 | 1064 | 99.6% | 11.7% |
| [773, 811) | 37 | 0 | 37 | 0.0% | 0.5% | 100.0% | 100.0% | 0.0% | 9165 | 1064 | 100.0% | 11.6% |

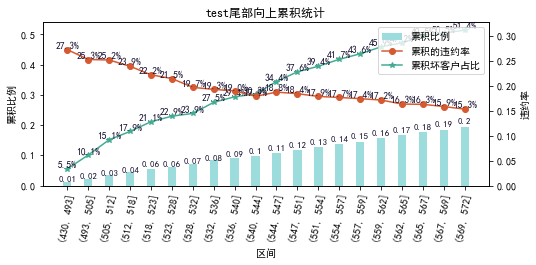


# 八、累积分数分布

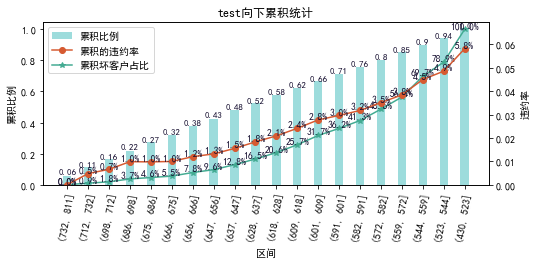
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

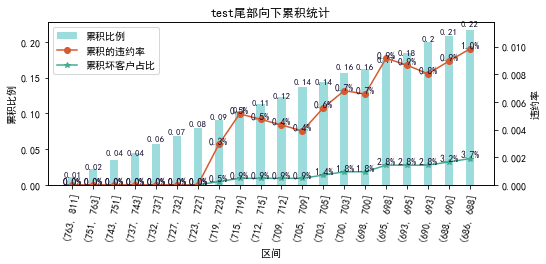
## 1. 向上累积分布





## 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 |
| [430.621, 468.9) | 0.2% | 0.2% | 0.1% | 0.0 | 0.002 |
| [468.9, 506.8) | 2.2% | 2.2% | 1.8% | 0.0 | 0.001 |
| [506.8, 544.7) | 8.0% | 7.4% | 8.4% | 0.0 | 0.0 |
| [544.7, 582.6) | 14.7% | 14.5% | 15.7% | 0.0 | 0.001 |
| [582.6, 620.5) | 19.6% | 19.0% | 22.6% | 0.0 | 0.004 |
| [620.5, 658.4) | 19.8% | 20.1% | 20.7% | 0.0 | 0.0 |
| [658.4, 696.3) | 18.4% | 19.5% | 17.5% | 0.001 | 0.0 |
| [696.3, 734.2) | 12.2% | 11.7% | 9.6% | 0.0 | 0.007 |
| [734.2, 772.1) | 4.3% | 4.8% | 3.4% | 0.0 | 0.002 |
| [772.1, 810.0) | 0.5% | 0.6% | 0.4% | 0.0 | 0.0 |
| sum | 100.0% | 100.0% | 100.0% | 0.002 | 0.018 |

# 十、策略建议