AKULAKU：id\_pdlnew\_17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 | 2020-05-09 |

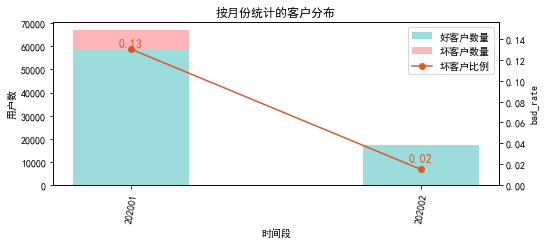
# 一、建模样本

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 |
|  | 84320 | 0.1068 | 202001 | 202002 | -- | -- |

表2 按月份统计分布表

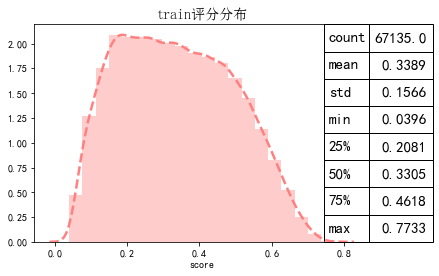
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 202001 | 58386.0 | 8749.0 | 67135 | 13.0% | 58386.0 | 8749.0 | 67135.0 | 13.0% | 0.0 | 0.0 | 0.0 | 0.0% |
| 202002 | 16926.0 | 259.0 | 17185 | 1.5% | 0.0 | 0.0 | 0.0 | 0.0% | 16926.0 | 259.0 | 17185.0 | 1.5% |

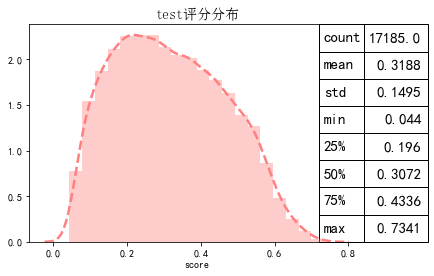


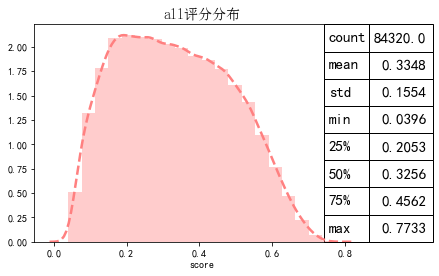
# 二、效果概况

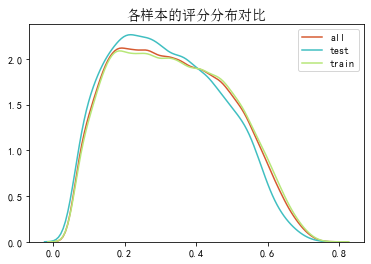
表3 模型效果概况表

|  |  |  |  |
| --- | --- | --- | --- |
|  | train | test | ALL |
| 样本量 | 67135 | 17185 | 84320 |
| 坏客户数量 | 8749 | 259 | 9008 |
| 坏客户比例 | 13.0% | 1.5% | 10.7% |
| KS | 0.356 | 0.35 | 0.357 |
| AUC | 0.26 | 0.275 | 0.258 |



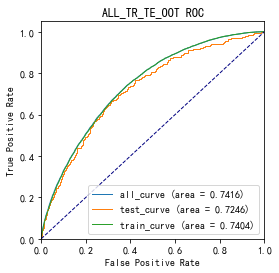
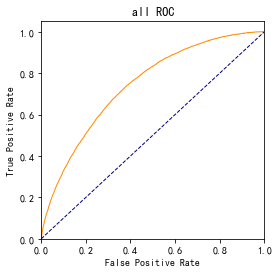
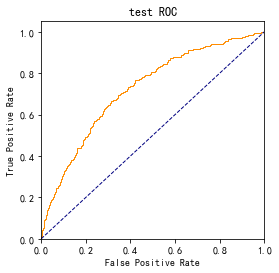
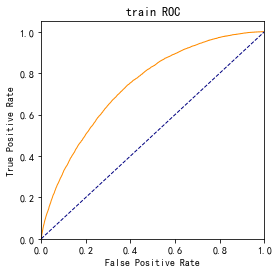






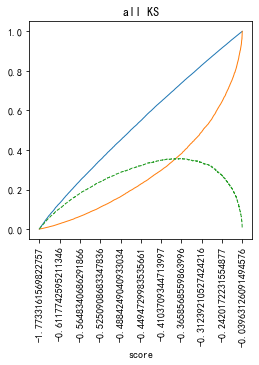
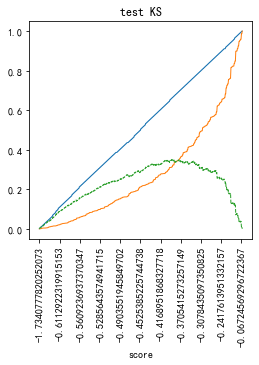
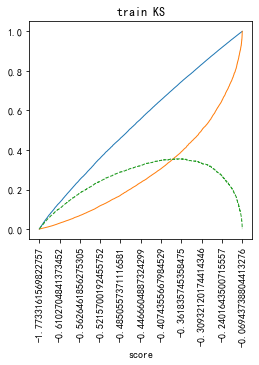
# 三、ROC曲线

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

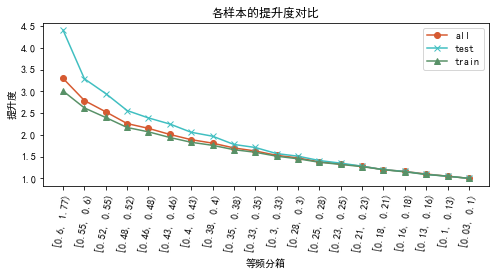
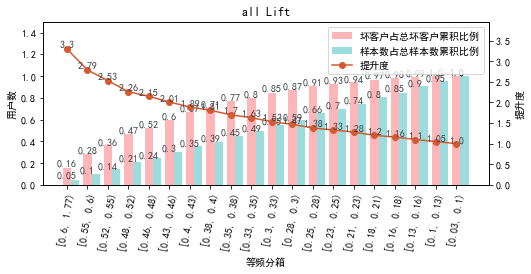
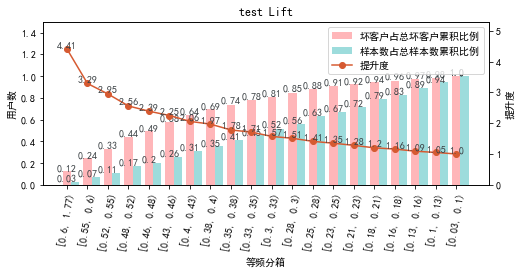
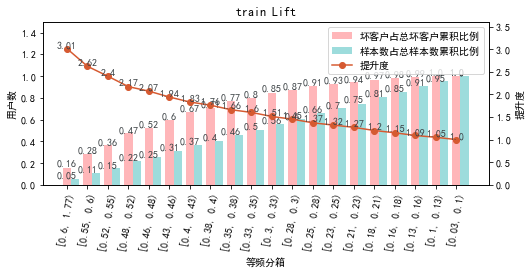
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

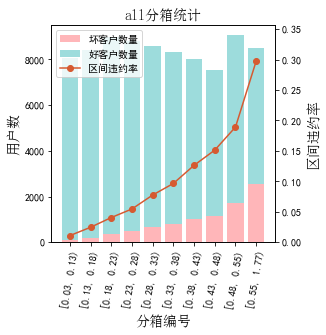
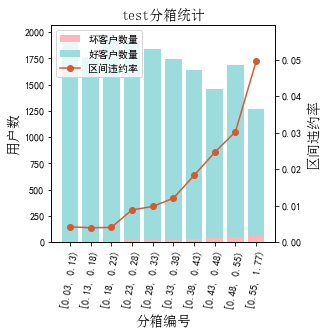
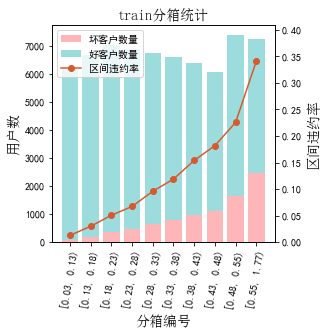
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



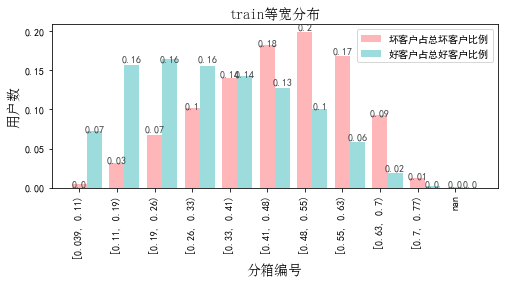
# 七、分数分布

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

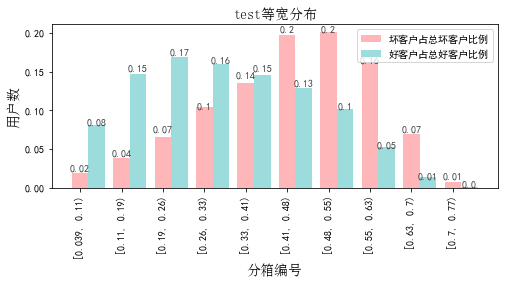
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [0.039, 0.11) | 4291 | 37 | 4254 | 0.4% | 7.3% | 0.4% | 7.3% | 0.9% | 4291 | 37 | 6.4% | 0.9% |
| [0.11, 0.19) | 9471 | 275 | 9196 | 3.1% | 15.8% | 3.6% | 23.0% | 2.9% | 13762 | 312 | 20.5% | 2.3% |
| [0.19, 0.26) | 10180 | 594 | 9586 | 6.8% | 16.4% | 10.4% | 39.5% | 5.8% | 23942 | 906 | 35.7% | 3.8% |
| [0.26, 0.33) | 9976 | 892 | 9084 | 10.2% | 15.6% | 20.6% | 55.0% | 8.9% | 33918 | 1798 | 50.5% | 5.3% |
| [0.33, 0.41) | 9558 | 1223 | 8335 | 14.0% | 14.3% | 34.5% | 69.3% | 12.8% | 43476 | 3021 | 64.8% | 6.9% |
| [0.41, 0.48) | 9033 | 1594 | 7439 | 18.2% | 12.7% | 52.7% | 82.0% | 17.6% | 52509 | 4615 | 78.2% | 8.8% |
| [0.48, 0.55) | 7651 | 1740 | 5911 | 19.9% | 10.1% | 72.6% | 92.2% | 22.7% | 60160 | 6355 | 89.6% | 10.6% |
| [0.55, 0.63) | 4853 | 1475 | 3378 | 16.9% | 5.8% | 89.5% | 97.9% | 30.4% | 65013 | 7830 | 96.8% | 12.0% |
| [0.63, 0.7) | 1910 | 808 | 1102 | 9.2% | 1.9% | 98.7% | 99.8% | 42.3% | 66923 | 8638 | 99.7% | 12.9% |
| [0.7, 0.77) | 211 | 110 | 101 | 1.3% | 0.2% | 100.0% | 100.0% | 52.1% | 67134 | 8748 | 100.0% | 13.0% |
| nan | 1 | 1 | 0 | 0.0% | 0.0% | 100.0% | 100.0% | 100.0% | 67135 | 8749 | 100.0% | 13.0% |



## 2.test

表5 等宽分箱数据分布表

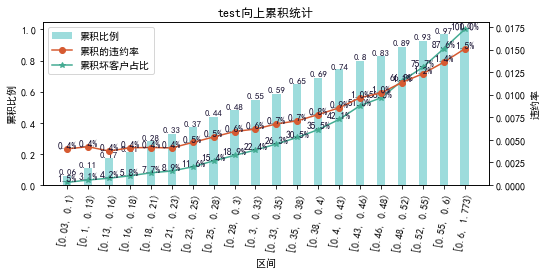
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [0.039, 0.11) | 1373 | 5 | 1368 | 1.9% | 8.1% | 1.9% | 8.1% | 0.4% | 1373 | 5 | 8.0% | 0.4% |
| [0.11, 0.19) | 2504 | 10 | 2494 | 3.9% | 14.7% | 5.8% | 22.8% | 0.4% | 3877 | 15 | 22.6% | 0.4% |
| [0.19, 0.26) | 2879 | 17 | 2862 | 6.6% | 16.9% | 12.4% | 39.7% | 0.6% | 6756 | 32 | 39.3% | 0.5% |
| [0.26, 0.33) | 2741 | 27 | 2714 | 10.4% | 16.0% | 22.8% | 55.8% | 1.0% | 9497 | 59 | 55.3% | 0.6% |
| [0.33, 0.41) | 2505 | 35 | 2470 | 13.5% | 14.6% | 36.3% | 70.4% | 1.4% | 12002 | 94 | 69.8% | 0.8% |
| [0.41, 0.48) | 2225 | 51 | 2174 | 19.7% | 12.8% | 56.0% | 83.2% | 2.3% | 14227 | 145 | 82.8% | 1.0% |
| [0.48, 0.55) | 1776 | 52 | 1724 | 20.1% | 10.2% | 76.1% | 93.4% | 2.9% | 16003 | 197 | 93.1% | 1.2% |
| [0.55, 0.63) | 920 | 42 | 878 | 16.2% | 5.2% | 92.3% | 98.6% | 4.6% | 16923 | 239 | 98.5% | 1.4% |
| [0.63, 0.7) | 242 | 18 | 224 | 6.9% | 1.3% | 99.2% | 99.9% | 7.4% | 17165 | 257 | 99.9% | 1.5% |
| [0.7, 0.77) | 20 | 2 | 18 | 0.8% | 0.1% | 100.0% | 100.0% | 10.0% | 17185 | 259 | 100.0% | 1.5% |

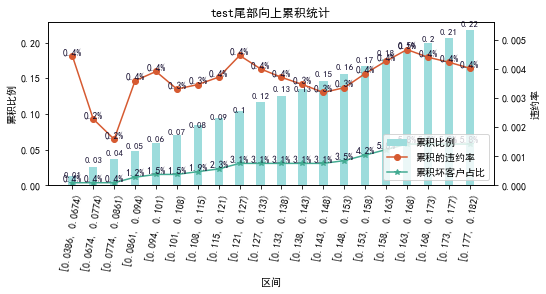


# 八、累积分数分布

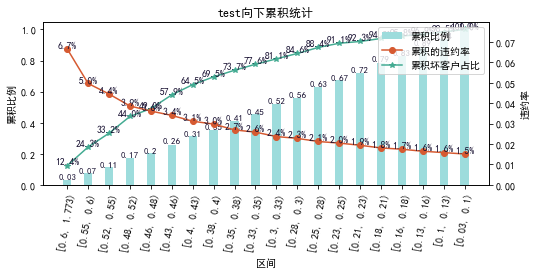
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

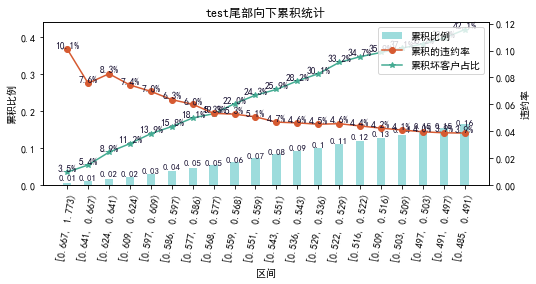
## 1. 向上累积分布





## 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 | train\_test\_psi |
| [0.0389, 0.113) | 6.4% | 8.0% | 0.004 |
| [0.113, 0.186) | 14.1% | 14.6% | 0.0 |
| [0.186, 0.26) | 15.2% | 16.8% | 0.002 |
| [0.26, 0.333) | 14.9% | 15.9% | 0.001 |
| [0.333, 0.406) | 14.2% | 14.6% | 0.0 |
| [0.406, 0.48) | 13.5% | 12.9% | 0.0 |
| [0.48, 0.553) | 11.4% | 10.3% | 0.001 |
| [0.553, 0.627) | 7.2% | 5.4% | 0.006 |
| [0.627, 0.7) | 2.8% | 1.4% | 0.01 |
| [0.7, 0.773) | 0.3% | 0.1% | 0.002 |
| sum | 100.0% | 100.0% | 0.025 |

# 十、策略建议