AKULAKU：id\_pdlold\_81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 | 2020-10-12 |

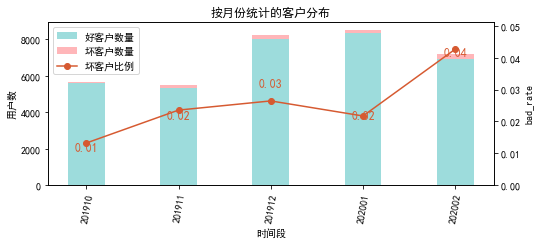
# 一、建模样本

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 |
|  | 35157 | 0.026 | 201910 | 202002 | -- | -- |

表2 按月份统计分布表

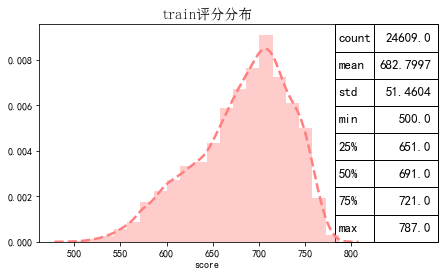
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 201910 | 5606 | 75 | 5681 | 1.3% | 3934 | 52 | 3986 | 1.3% | 1672 | 23 | 1695 | 1.4% |
| 201911 | 5351 | 129 | 5480 | 2.4% | 3757 | 76 | 3833 | 2.0% | 1594 | 53 | 1647 | 3.2% |
| 201912 | 8033 | 218 | 8251 | 2.6% | 5601 | 160 | 5761 | 2.8% | 2432 | 58 | 2490 | 2.3% |
| 202001 | 8343 | 185 | 8528 | 2.2% | 5877 | 129 | 6006 | 2.1% | 2466 | 56 | 2522 | 2.2% |
| 202002 | 6909 | 308 | 7217 | 4.3% | 4800 | 223 | 5023 | 4.4% | 2109 | 85 | 2194 | 3.9% |

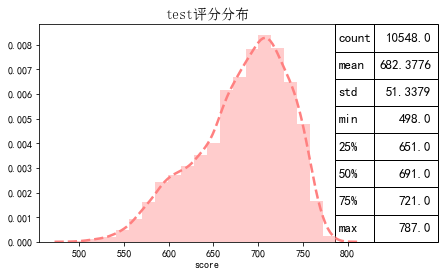


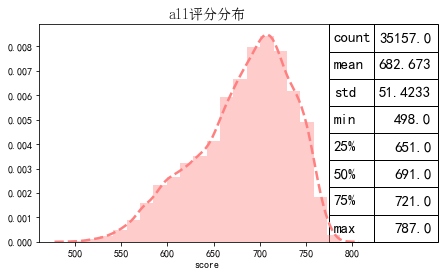
# 二、效果概况

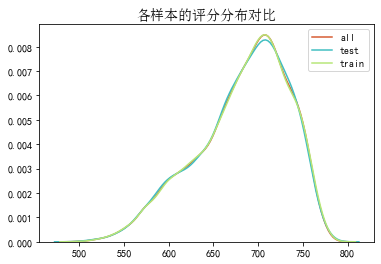
表3 模型效果概况表

|  |  |  |  |
| --- | --- | --- | --- |
|  | train | test | ALL |
| 样本量 | 24609 | 10548 | 35157 |
| 坏客户数量 | 640 | 275 | 915 |
| 坏客户比例 | 2.6% | 2.6% | 2.6% |
| KS | 0.388 | 0.387 | 0.388 |
| AUC | 0.752 | 0.759 | 0.754 |



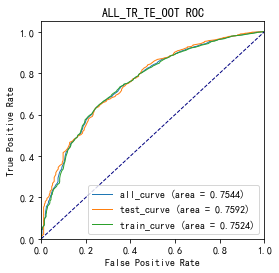
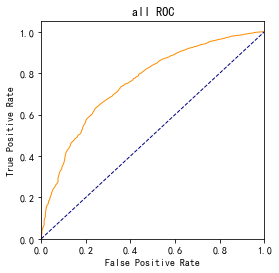
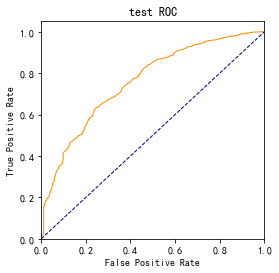
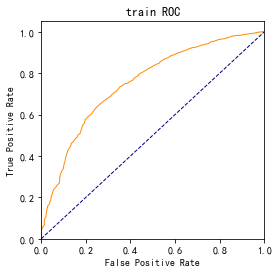






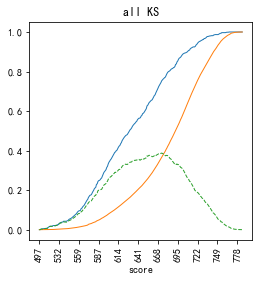
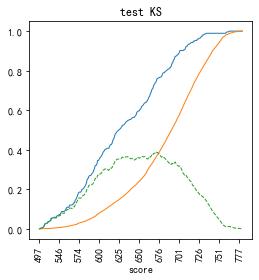
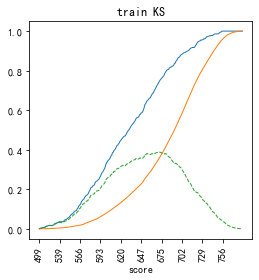
# 三、ROC曲线

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

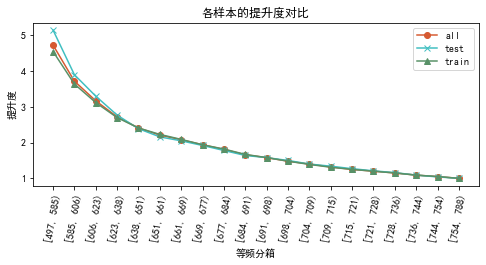
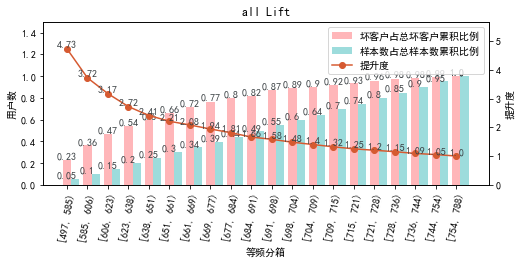
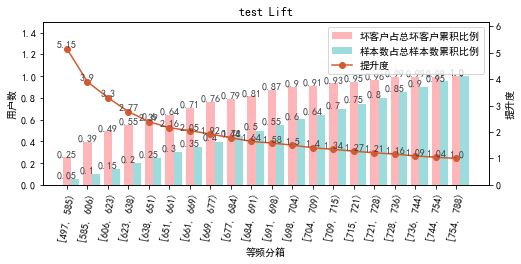
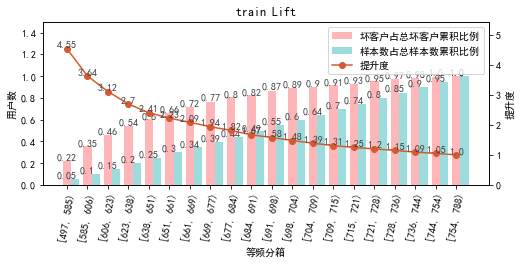
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

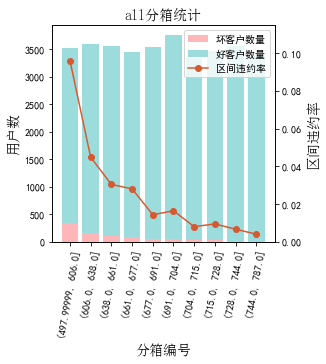
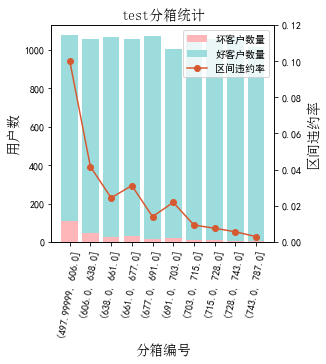
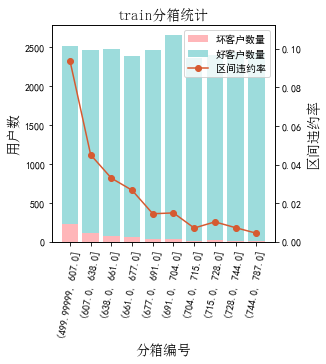
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



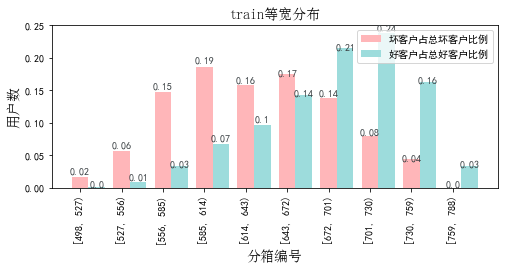
# 七、分数分布

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

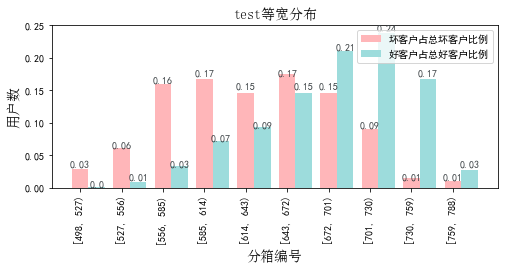
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [498, 527) | 46 | 11 | 35 | 1.7% | 0.1% | 1.7% | 0.1% | 23.9% | 46 | 11 | 0.2% | 23.9% |
| [527, 556) | 266 | 36 | 230 | 5.6% | 1.0% | 7.3% | 1.1% | 13.5% | 312 | 47 | 1.3% | 15.1% |
| [556, 585) | 880 | 94 | 786 | 14.7% | 3.3% | 22.0% | 4.4% | 10.7% | 1192 | 141 | 4.8% | 11.8% |
| [585, 614) | 1745 | 119 | 1626 | 18.6% | 6.8% | 40.6% | 11.2% | 6.8% | 2937 | 260 | 11.9% | 8.9% |
| [614, 643) | 2412 | 101 | 2311 | 15.8% | 9.6% | 56.4% | 20.8% | 4.2% | 5349 | 361 | 21.7% | 6.7% |
| [643, 672) | 3539 | 112 | 3427 | 17.5% | 14.3% | 73.9% | 35.1% | 3.2% | 8888 | 473 | 36.1% | 5.3% |
| [672, 701) | 5226 | 88 | 5138 | 13.8% | 21.4% | 87.7% | 56.5% | 1.7% | 14114 | 561 | 57.4% | 4.0% |
| [701, 730) | 5777 | 51 | 5726 | 8.0% | 23.9% | 95.6% | 80.4% | 0.9% | 19891 | 612 | 80.8% | 3.1% |
| [730, 759) | 3923 | 28 | 3895 | 4.4% | 16.3% | 100.0% | 96.7% | 0.7% | 23814 | 640 | 96.8% | 2.7% |
| [759, 788) | 795 | 0 | 795 | 0.0% | 3.3% | 100.0% | 100.0% | 0.0% | 24609 | 640 | 100.0% | 2.6% |



## 2.test

表5 等宽分箱数据分布表

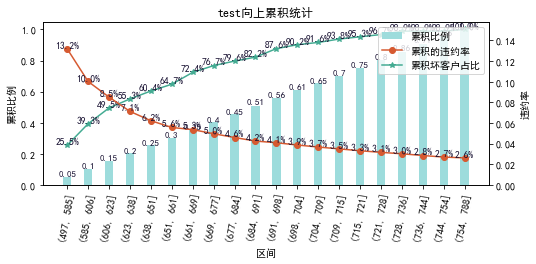
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [498, 527) | 23 | 8 | 15 | 2.9% | 0.1% | 2.9% | 0.1% | 34.8% | 23 | 8 | 0.2% | 34.8% |
| [527, 556) | 104 | 17 | 87 | 6.2% | 0.8% | 9.1% | 1.0% | 16.3% | 127 | 25 | 1.2% | 19.7% |
| [556, 585) | 387 | 44 | 343 | 16.0% | 3.3% | 25.1% | 4.3% | 11.4% | 514 | 69 | 4.9% | 13.4% |
| [585, 614) | 779 | 46 | 733 | 16.7% | 7.1% | 41.8% | 11.5% | 5.9% | 1293 | 115 | 12.3% | 8.9% |
| [614, 643) | 1005 | 40 | 965 | 14.5% | 9.4% | 56.4% | 20.9% | 4.0% | 2298 | 155 | 21.8% | 6.7% |
| [643, 672) | 1554 | 48 | 1506 | 17.5% | 14.7% | 73.8% | 35.5% | 3.1% | 3852 | 203 | 36.5% | 5.3% |
| [672, 701) | 2208 | 40 | 2168 | 14.5% | 21.1% | 88.4% | 56.6% | 1.8% | 6060 | 243 | 57.5% | 4.0% |
| [701, 730) | 2480 | 25 | 2455 | 9.1% | 23.9% | 97.5% | 80.5% | 1.0% | 8540 | 268 | 81.0% | 3.1% |
| [730, 759) | 1719 | 4 | 1715 | 1.5% | 16.7% | 98.9% | 97.2% | 0.2% | 10259 | 272 | 97.3% | 2.7% |
| [759, 788) | 289 | 3 | 286 | 1.1% | 2.8% | 100.0% | 100.0% | 1.0% | 10548 | 275 | 100.0% | 2.6% |

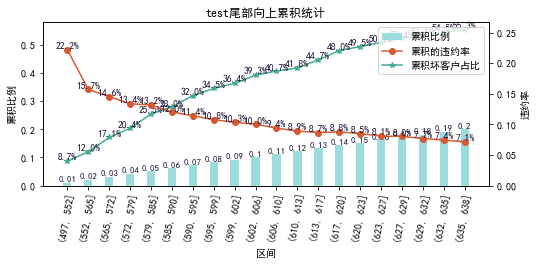


# 八、累积分数分布

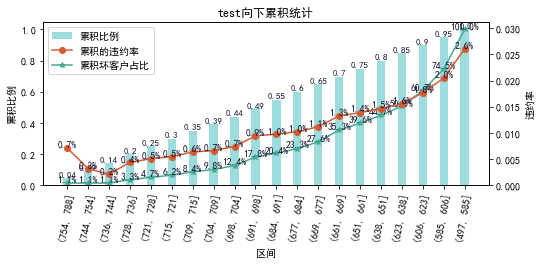
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

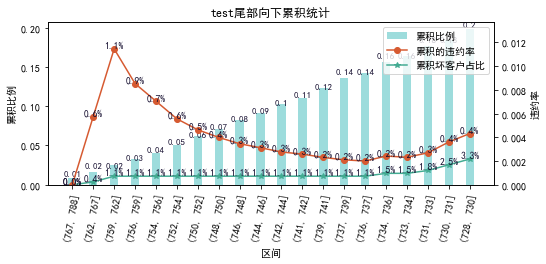
## 1. 向上累积分布





## 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 | train\_test\_psi |
| [497.711, 526.9) | 0.2% | 0.2% | 0.0 |
| [526.9, 555.8) | 1.1% | 1.0% | 0.0 |
| [555.8, 584.7) | 3.6% | 3.7% | 0.0 |
| [584.7, 613.6) | 7.1% | 7.4% | 0.0 |
| [613.6, 642.5) | 9.8% | 9.5% | 0.0 |
| [642.5, 671.4) | 14.4% | 14.7% | 0.0 |
| [671.4, 700.3) | 21.2% | 20.9% | 0.0 |
| [700.3, 729.2) | 23.5% | 23.5% | 0.0 |
| [729.2, 758.1) | 15.9% | 16.3% | 0.0 |
| [758.1, 787.0) | 3.2% | 2.7% | 0.001 |
| sum | 100.0% | 100.0% | 0.001 |

# 十、策略建议