AKULAKU：idpdlold73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 | 2020-09-25 |

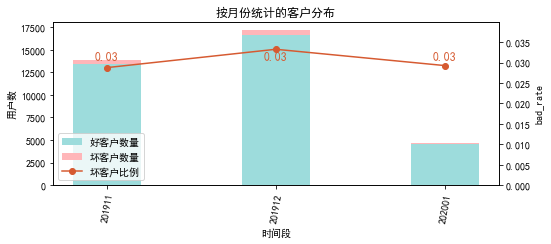
# 一、建模样本

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 |
|  | 35783 | 0.031 | 201911 | 202001 | -- | -- |

表2 按月份统计分布表

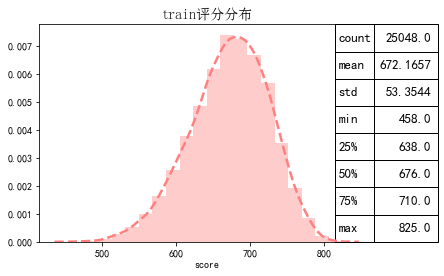
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 201911 | 13471 | 399 | 13870 | 2.9% | 9437 | 277 | 9714 | 2.9% | 4034 | 122 | 4156 | 2.9% |
| 201912 | 16626 | 572 | 17198 | 3.3% | 11640 | 397 | 12037 | 3.3% | 4986 | 175 | 5161 | 3.4% |
| 202001 | 4577 | 138 | 4715 | 2.9% | 3195 | 102 | 3297 | 3.1% | 1382 | 36 | 1418 | 2.5% |

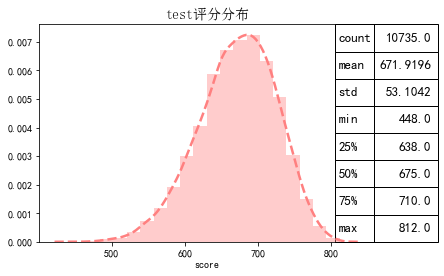


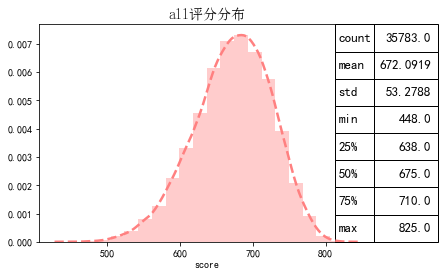
# 二、效果概况

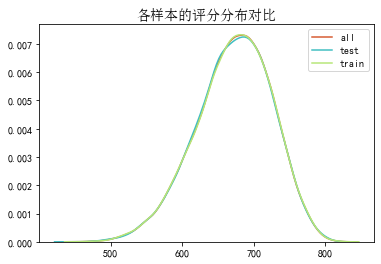
表3 模型效果概况表

|  |  |  |  |
| --- | --- | --- | --- |
|  | train | test | ALL |
| 样本量 | 25048 | 10735 | 35783 |
| 坏客户数量 | 776 | 333 | 1109 |
| 坏客户比例 | 3.1% | 3.1% | 3.1% |
| KS | 0.427 | 0.397 | 0.417 |
| AUC | 0.787 | 0.76 | 0.779 |



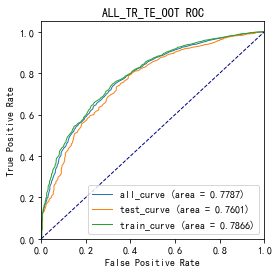
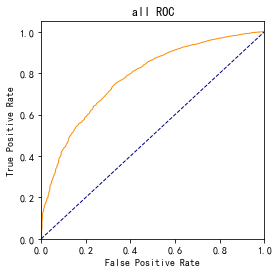
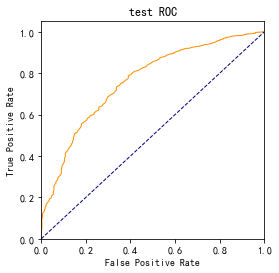
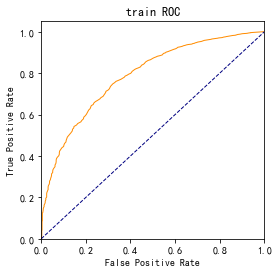






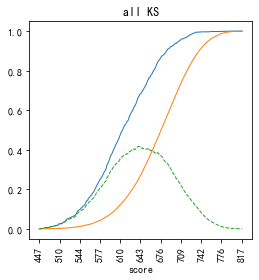
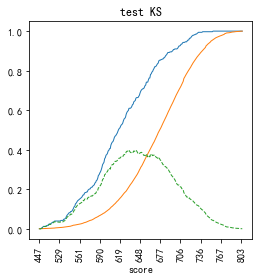
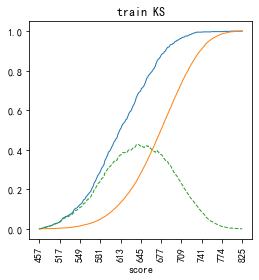
# 三、ROC曲线

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

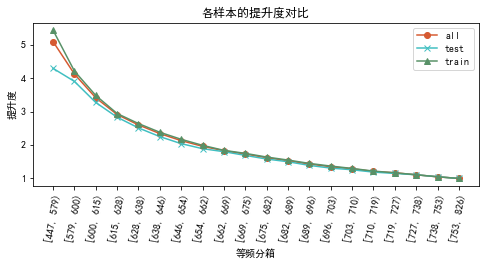
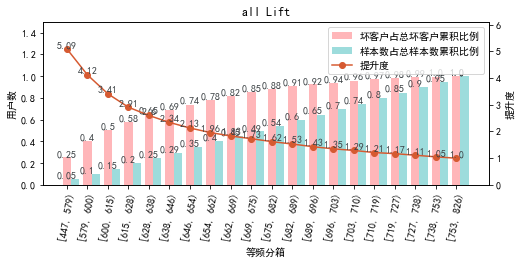
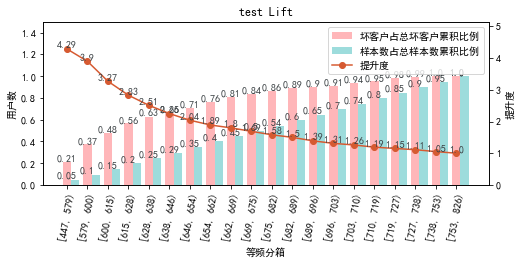
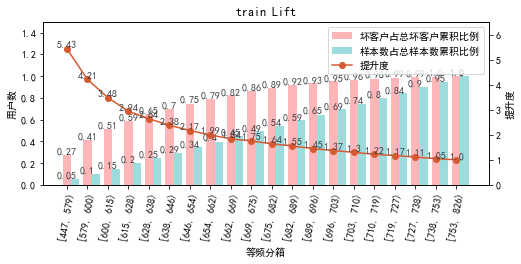
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

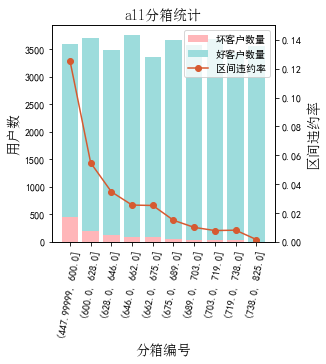
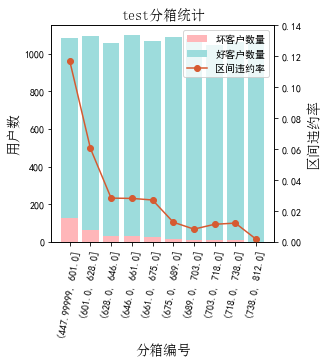
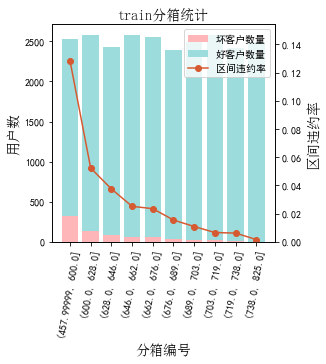
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



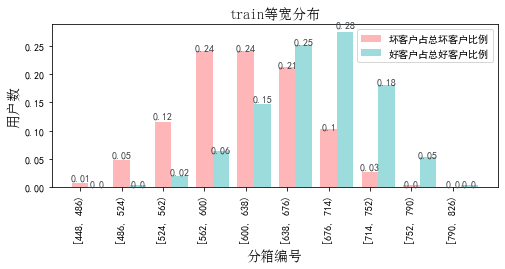
# 七、分数分布

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

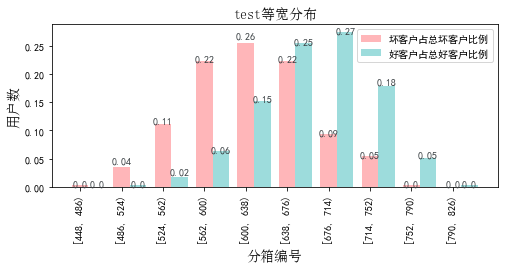
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [448, 486) | 12 | 6 | 6 | 0.8% | 0.0% | 0.8% | 0.0% | 50.0% | 12 | 6 | 0.0% | 50.0% |
| [486, 524) | 128 | 37 | 91 | 4.8% | 0.4% | 5.5% | 0.4% | 28.9% | 140 | 43 | 0.6% | 30.7% |
| [524, 562) | 568 | 90 | 478 | 11.6% | 2.0% | 17.1% | 2.4% | 15.8% | 708 | 133 | 2.8% | 18.8% |
| [562, 600) | 1745 | 187 | 1558 | 24.1% | 6.4% | 41.2% | 8.8% | 10.7% | 2453 | 320 | 9.8% | 13.0% |
| [600, 638) | 3749 | 187 | 3562 | 24.1% | 14.7% | 65.3% | 23.5% | 5.0% | 6202 | 507 | 24.8% | 8.2% |
| [638, 676) | 6291 | 165 | 6126 | 21.3% | 25.2% | 86.6% | 48.7% | 2.6% | 12493 | 672 | 49.9% | 5.4% |
| [676, 714) | 6755 | 80 | 6675 | 10.3% | 27.5% | 96.9% | 76.2% | 1.2% | 19248 | 752 | 76.8% | 3.9% |
| [714, 752) | 4407 | 21 | 4386 | 2.7% | 18.1% | 99.6% | 94.3% | 0.5% | 23655 | 773 | 94.4% | 3.3% |
| [752, 790) | 1296 | 3 | 1293 | 0.4% | 5.3% | 100.0% | 99.6% | 0.2% | 24951 | 776 | 99.6% | 3.1% |
| [790, 826) | 97 | 0 | 97 | 0.0% | 0.4% | 100.0% | 100.0% | 0.0% | 25048 | 776 | 100.0% | 3.1% |



## 2.test

表5 等宽分箱数据分布表

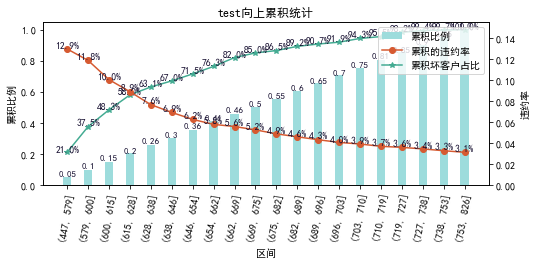
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [448, 486) | 5 | 1 | 4 | 0.3% | 0.0% | 0.3% | 0.0% | 20.0% | 5 | 1 | 0.0% | 20.0% |
| [486, 524) | 55 | 12 | 43 | 3.6% | 0.4% | 3.9% | 0.5% | 21.8% | 60 | 13 | 0.6% | 21.7% |
| [524, 562) | 231 | 37 | 194 | 11.1% | 1.9% | 15.0% | 2.3% | 16.0% | 291 | 50 | 2.7% | 17.2% |
| [562, 600) | 733 | 74 | 659 | 22.2% | 6.3% | 37.2% | 8.7% | 10.1% | 1024 | 124 | 9.5% | 12.1% |
| [600, 638) | 1658 | 85 | 1573 | 25.5% | 15.1% | 62.8% | 23.8% | 5.1% | 2682 | 209 | 25.0% | 7.8% |
| [638, 676) | 2717 | 74 | 2643 | 22.2% | 25.4% | 85.0% | 49.2% | 2.7% | 5399 | 283 | 50.3% | 5.2% |
| [676, 714) | 2885 | 31 | 2854 | 9.3% | 27.4% | 94.3% | 76.6% | 1.1% | 8284 | 314 | 77.2% | 3.8% |
| [714, 752) | 1873 | 18 | 1855 | 5.4% | 17.8% | 99.7% | 94.5% | 1.0% | 10157 | 332 | 94.6% | 3.3% |
| [752, 790) | 537 | 1 | 536 | 0.3% | 5.2% | 100.0% | 99.6% | 0.2% | 10694 | 333 | 99.6% | 3.1% |
| [790, 826) | 41 | 0 | 41 | 0.0% | 0.4% | 100.0% | 100.0% | 0.0% | 10735 | 333 | 100.0% | 3.1% |

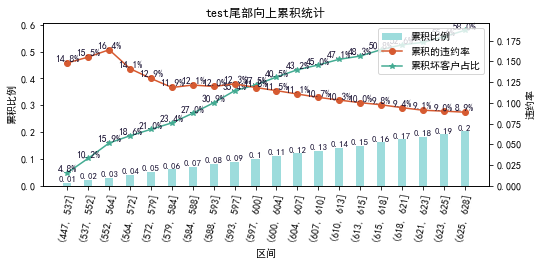


# 八、累积分数分布

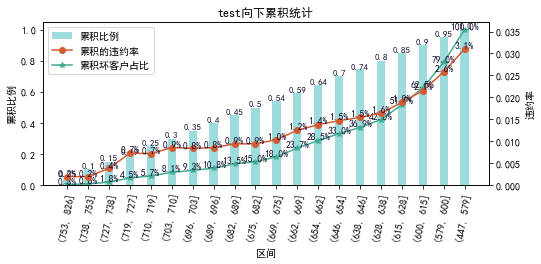
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

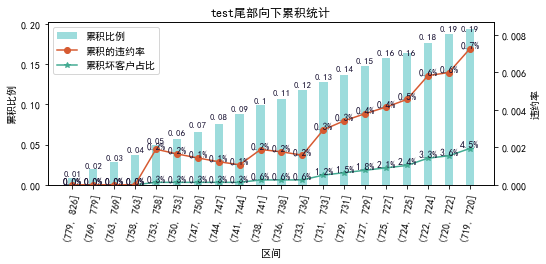
## 1. 向上累积分布





## 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 |
| [447.623, 485.7) | 0.0% | 0.0% | 0.0 |
| [485.7, 523.4) | 0.5% | 0.5% | 0.0 |
| [523.4, 561.1) | 2.3% | 2.2% | 0.0 |
| [561.1, 598.8) | 6.7% | 6.6% | 0.0 |
| [598.8, 636.5) | 14.8% | 15.2% | 0.0 |
| [636.5, 674.2) | 24.8% | 25.2% | 0.0 |
| [674.2, 711.9) | 26.6% | 26.3% | 0.0 |
| [711.9, 749.6) | 18.2% | 18.2% | 0.0 |
| [749.6, 787.3) | 5.6% | 5.5% | 0.0 |
| [787.3, 825.0) | 0.4% | 0.4% | 0.0 |
| sum | 100.0% | 100.0% | 0.0 |

# 十、策略建议