AKULAKU：idpdlold72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 |

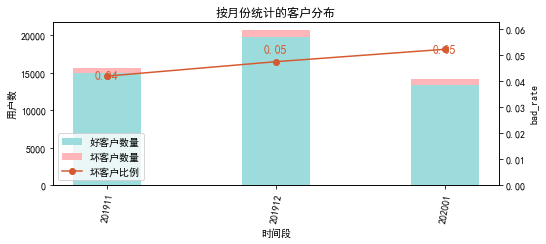
# 一、建模样本

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 |
|  | 50448 | 0.0471 | 201911 | 202001 | -- | -- |

表2 按月份统计分布表

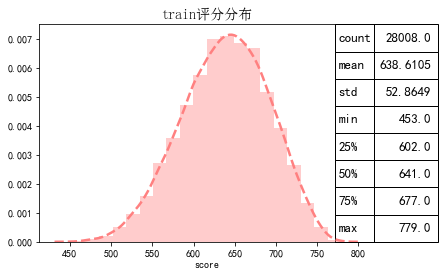
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 201911 | 14936 | 653 | 15589 | 4.2% | 13381.0 | 599.0 | 13980.0 | 4.3% | 1555 | 54 | 1609 | 3.4% |
| 201912 | 19758 | 984 | 20742 | 4.7% | 13296.0 | 732.0 | 14028.0 | 5.2% | 6462 | 252 | 6714 | 3.8% |
| 202001 | 13380 | 737 | 14117 | 5.2% | 0.0 | 0.0 | 0.0 | 0.0% | 13380 | 737 | 14117 | 5.2% |

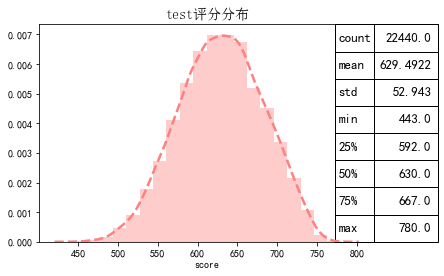


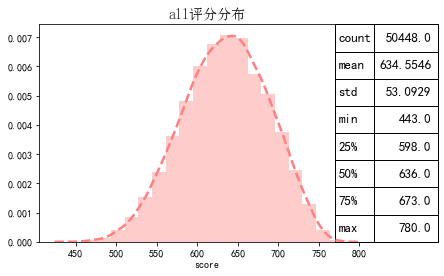
# 二、效果概况

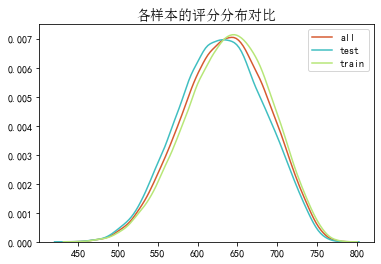
表3 模型效果概况表

|  |  |  |  |
| --- | --- | --- | --- |
|  | train | test | ALL |
| 样本量 | 28008 | 22440 | 50448 |
| 坏客户数量 | 1331 | 1043 | 2374 |
| 坏客户比例 | 4.8% | 4.6% | 4.7% |
| KS | 0.423 | 0.379 | 0.402 |
| AUC | 0.779 | 0.751 | 0.766 |



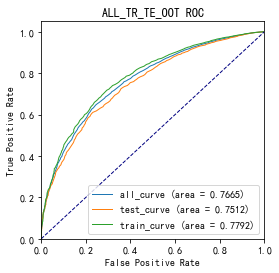
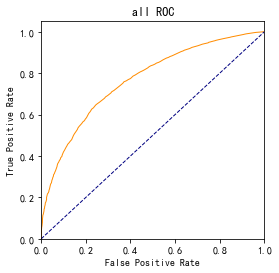
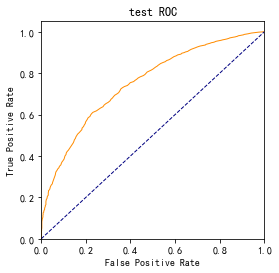
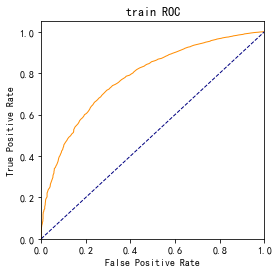






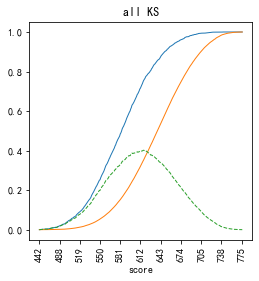
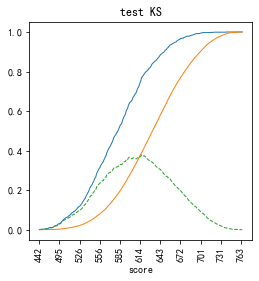
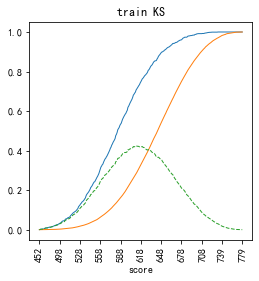
# 三、ROC曲线

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

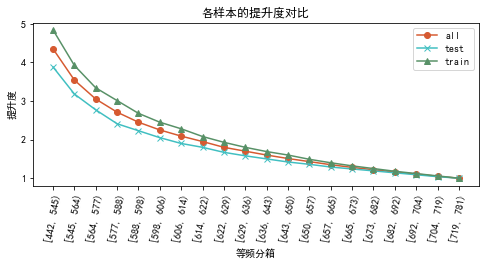
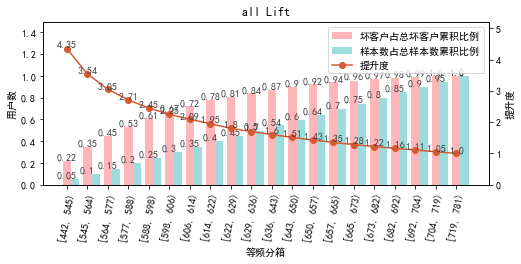
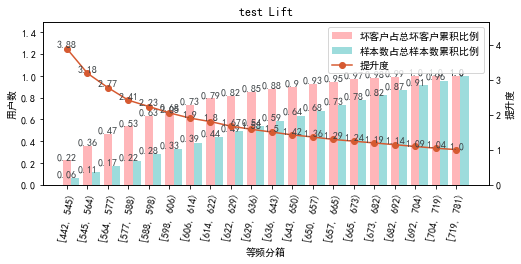
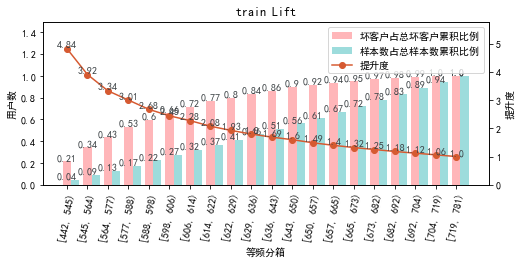
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

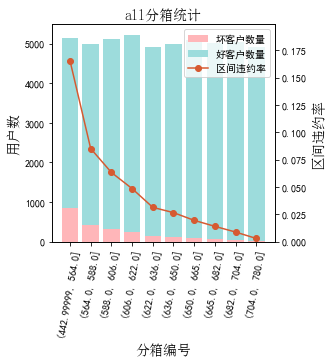
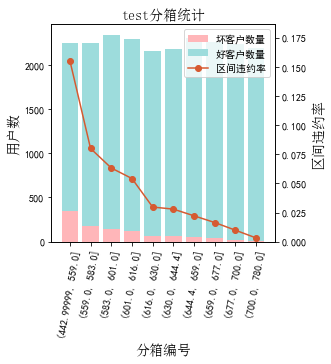
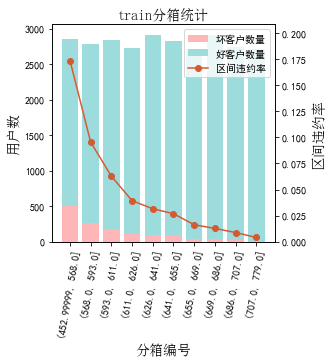
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



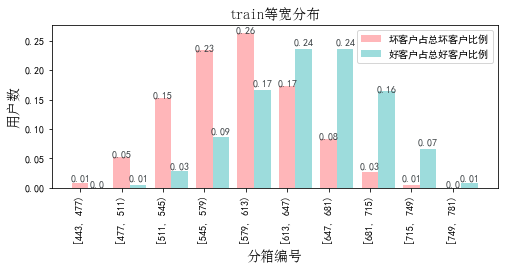
# 七、分数分布

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

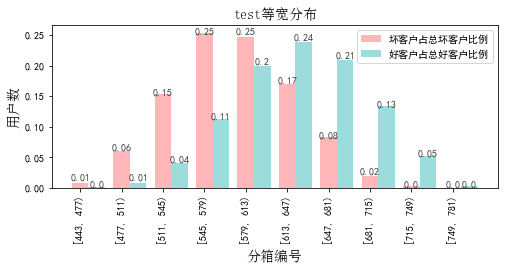
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [443, 477) | 26 | 11 | 15 | 0.8% | 0.1% | 0.8% | 0.1% | 42.3% | 26 | 11 | 0.1% | 42.3% |
| [477, 511) | 217 | 69 | 148 | 5.2% | 0.6% | 6.0% | 0.6% | 31.8% | 243 | 80 | 0.9% | 32.9% |
| [511, 545) | 992 | 204 | 788 | 15.3% | 3.0% | 21.3% | 3.6% | 20.6% | 1235 | 284 | 4.4% | 23.0% |
| [545, 579) | 2609 | 312 | 2297 | 23.4% | 8.6% | 44.8% | 12.2% | 12.0% | 3844 | 596 | 13.7% | 15.5% |
| [579, 613) | 4808 | 351 | 4457 | 26.4% | 16.7% | 71.1% | 28.9% | 7.3% | 8652 | 947 | 30.9% | 10.9% |
| [613, 647) | 6528 | 231 | 6297 | 17.4% | 23.6% | 88.5% | 52.5% | 3.5% | 15180 | 1178 | 54.2% | 7.8% |
| [647, 681) | 6409 | 110 | 6299 | 8.3% | 23.6% | 96.8% | 76.1% | 1.7% | 21589 | 1288 | 77.1% | 6.0% |
| [681, 715) | 4421 | 36 | 4385 | 2.7% | 16.4% | 99.5% | 92.5% | 0.8% | 26010 | 1324 | 92.9% | 5.1% |
| [715, 749) | 1781 | 7 | 1774 | 0.5% | 6.6% | 100.0% | 99.2% | 0.4% | 27791 | 1331 | 99.2% | 4.8% |
| [749, 781) | 217 | 0 | 217 | 0.0% | 0.8% | 100.0% | 100.0% | 0.0% | 28008 | 1331 | 100.0% | 4.8% |



## 2.test

表5 等宽分箱数据分布表

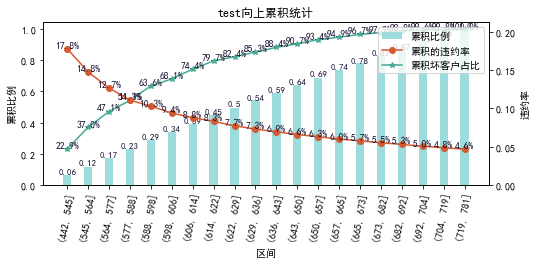
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [443, 477) | 25 | 9 | 16 | 0.9% | 0.1% | 0.9% | 0.1% | 36.0% | 25 | 9 | 0.1% | 36.0% |
| [477, 511) | 234 | 63 | 171 | 6.0% | 0.8% | 6.9% | 0.9% | 26.9% | 259 | 72 | 1.2% | 27.8% |
| [511, 545) | 1027 | 160 | 867 | 15.3% | 4.1% | 22.2% | 4.9% | 15.6% | 1286 | 232 | 5.7% | 18.0% |
| [545, 579) | 2683 | 265 | 2418 | 25.4% | 11.3% | 47.7% | 16.2% | 9.9% | 3969 | 497 | 17.7% | 12.5% |
| [579, 613) | 4531 | 257 | 4274 | 24.6% | 20.0% | 72.3% | 36.2% | 5.7% | 8500 | 754 | 37.9% | 8.9% |
| [613, 647) | 5271 | 178 | 5093 | 17.1% | 23.8% | 89.4% | 60.0% | 3.4% | 13771 | 932 | 61.4% | 6.8% |
| [647, 681) | 4581 | 87 | 4494 | 8.3% | 21.0% | 97.7% | 81.0% | 1.9% | 18352 | 1019 | 81.8% | 5.6% |
| [681, 715) | 2875 | 21 | 2854 | 2.0% | 13.3% | 99.7% | 94.3% | 0.7% | 21227 | 1040 | 94.6% | 4.9% |
| [715, 749) | 1136 | 3 | 1133 | 0.3% | 5.3% | 100.0% | 99.6% | 0.3% | 22363 | 1043 | 99.7% | 4.7% |
| [749, 781) | 77 | 0 | 77 | 0.0% | 0.4% | 100.0% | 100.0% | 0.0% | 22440 | 1043 | 100.0% | 4.6% |

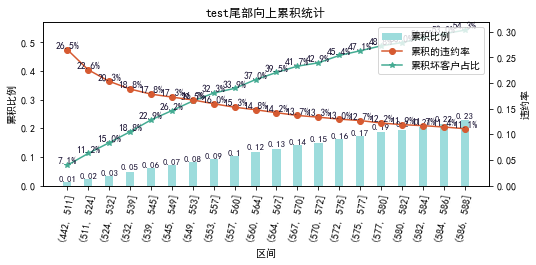


# 八、累积分数分布

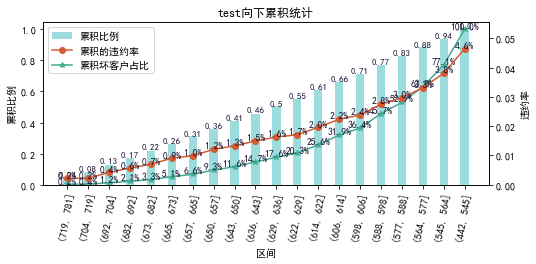
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

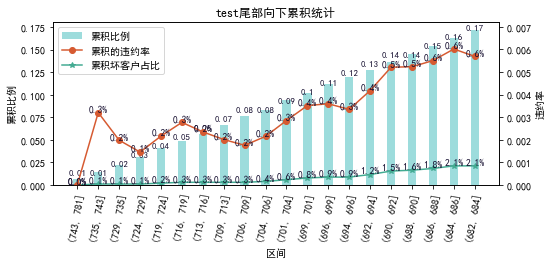
## 1. 向上累积分布





## 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 |
| [442.663, 476.7) | 0.1% | 0.1% | 0.0 |
| [476.7, 510.4) | 0.8% | 1.0% | 0.001 |
| [510.4, 544.1) | 3.5% | 4.6% | 0.003 |
| [544.1, 577.8) | 8.9% | 11.5% | 0.007 |
| [577.8, 611.5) | 16.9% | 19.9% | 0.005 |
| [611.5, 645.2) | 23.1% | 23.6% | 0.0 |
| [645.2, 678.9) | 22.5% | 20.2% | 0.002 |
| [678.9, 712.6) | 16.2% | 13.3% | 0.006 |
| [712.6, 746.3) | 6.9% | 5.5% | 0.004 |
| [746.3, 780.0) | 1.0% | 0.4% | 0.004 |
| sum | 100.0% | 100.0% | 0.031 |

# 十、策略建议