AKULAKU：消费分期老客payoffbili6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 | 2020-05-08 |

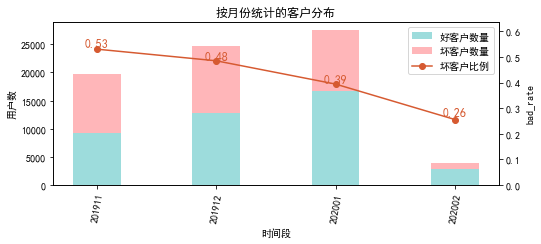
# 一、建模样本

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 |
|  | 75845 | 0.4519 | 201911 | 202002 | -- | -- |

表2 按月份统计分布表

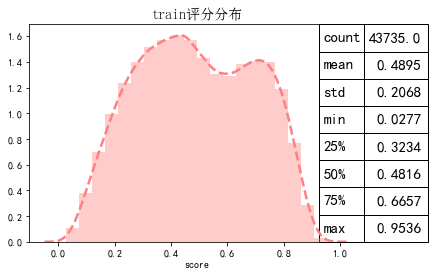
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 201911 | 9241 | 10434 | 19675 | 53.0% | 9241.0 | 10433.0 | 19674.0 | 53.0% | 0 | 1 | 1 | 100.0% |
| 201912 | 12721 | 11975 | 24696 | 48.5% | 11251.0 | 10147.0 | 21398.0 | 47.4% | 1470 | 1828 | 3298 | 55.4% |
| 202001 | 16716 | 10870 | 27586 | 39.4% | 1833.0 | 830.0 | 2663.0 | 31.2% | 14883 | 10040 | 24923 | 40.3% |
| 202002 | 2894 | 994 | 3888 | 25.6% | 0.0 | 0.0 | 0.0 | 0.0% | 2894 | 994 | 3888 | 25.6% |

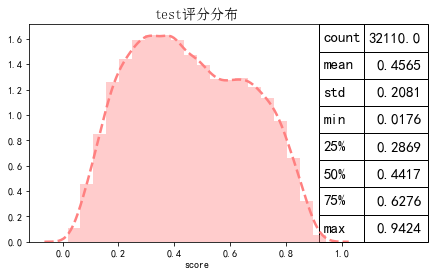


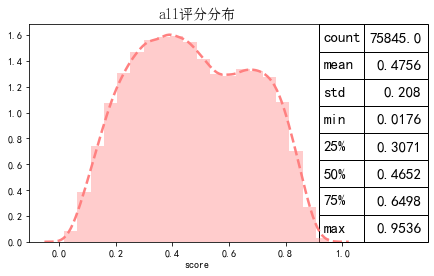
# 二、效果概况

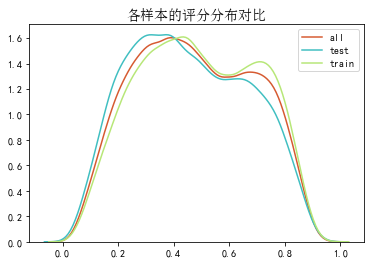
表3 模型效果概况表

|  |  |  |  |
| --- | --- | --- | --- |
|  | train | test | ALL |
| 样本量 | 43735 | 32110 | 75845 |
| 坏客户数量 | 21410 | 12863 | 34273 |
| 坏客户比例 | 49.0% | 40.1% | 45.2% |
| KS | 0.383 | 0.343 | 0.369 |
| AUC | 0.247 | 0.269 | 0.255 |



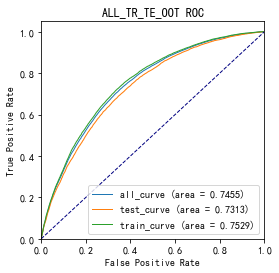
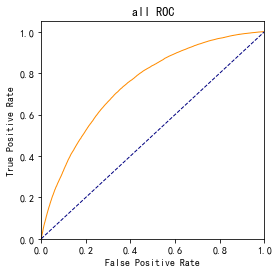
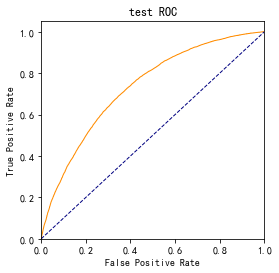
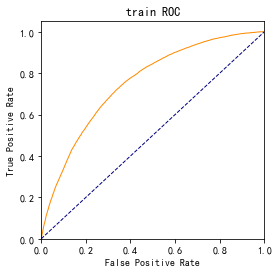






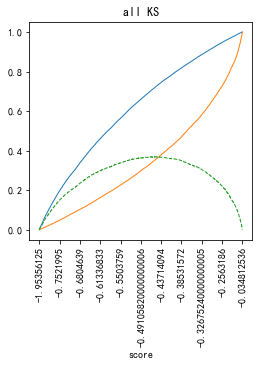
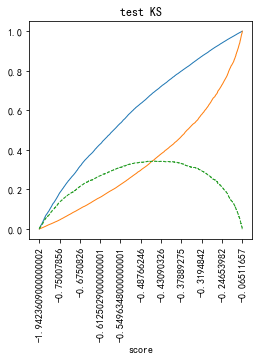
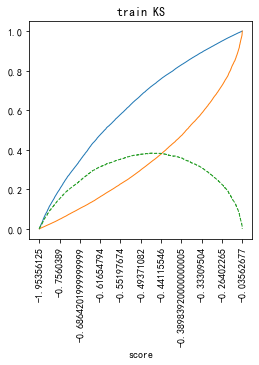
# 三、ROC曲线

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

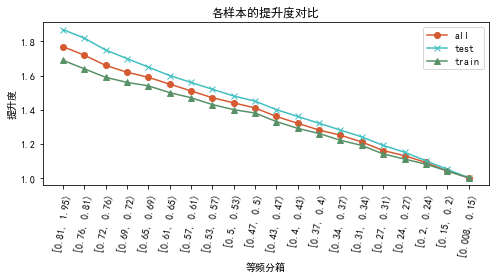
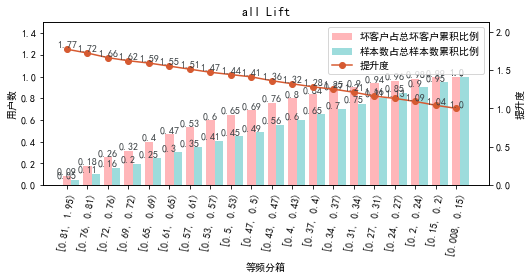
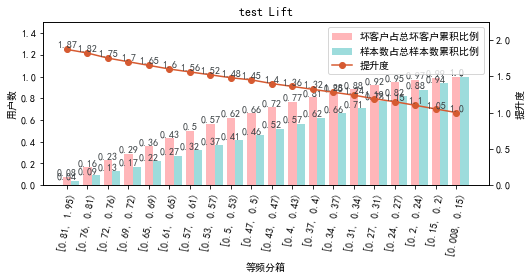
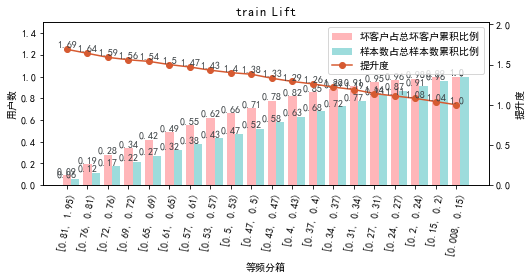
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

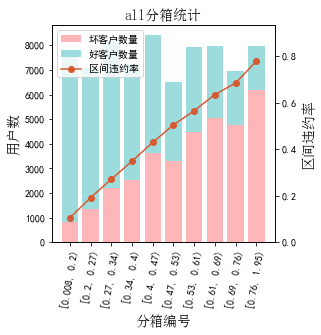
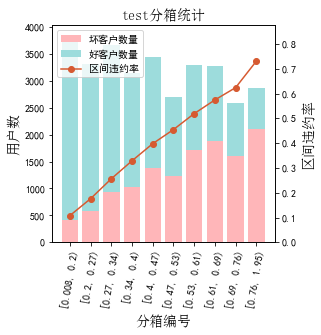
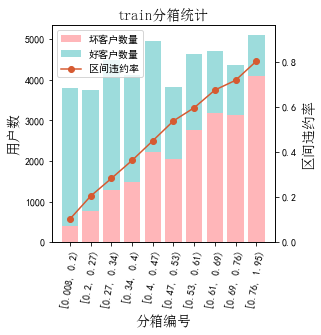
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



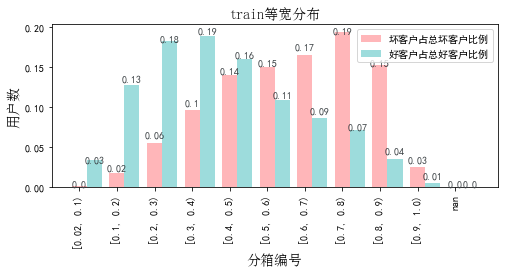
# 七、分数分布

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

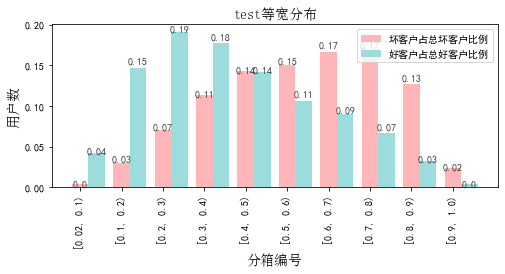
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [0.02, 0.1) | 791 | 40 | 751 | 0.2% | 3.4% | 0.2% | 3.4% | 5.1% | 791 | 40 | 1.8% | 5.1% |
| [0.1, 0.2) | 3234 | 394 | 2840 | 1.8% | 12.7% | 2.0% | 16.1% | 12.2% | 4025 | 434 | 9.2% | 10.8% |
| [0.2, 0.3) | 5281 | 1197 | 4084 | 5.6% | 18.3% | 7.6% | 34.4% | 22.7% | 9306 | 1631 | 21.3% | 17.5% |
| [0.3, 0.4) | 6276 | 2070 | 4206 | 9.7% | 18.8% | 17.3% | 53.2% | 33.0% | 15582 | 3701 | 35.6% | 23.8% |
| [0.4, 0.5) | 6561 | 3000 | 3561 | 14.0% | 16.0% | 31.3% | 69.2% | 45.7% | 22143 | 6701 | 50.6% | 30.3% |
| [0.5, 0.6) | 5639 | 3208 | 2431 | 15.0% | 10.9% | 46.3% | 80.1% | 56.9% | 27782 | 9909 | 63.5% | 35.7% |
| [0.6, 0.7) | 5461 | 3535 | 1926 | 16.5% | 8.6% | 62.8% | 88.7% | 64.7% | 33243 | 13444 | 76.0% | 40.4% |
| [0.7, 0.8) | 5753 | 4142 | 1611 | 19.3% | 7.2% | 82.1% | 95.9% | 72.0% | 38996 | 17586 | 89.2% | 45.1% |
| [0.8, 0.9) | 4075 | 3273 | 802 | 15.3% | 3.6% | 97.4% | 99.5% | 80.3% | 43071 | 20859 | 98.5% | 48.4% |
| [0.9, 1.0) | 663 | 550 | 113 | 2.6% | 0.5% | 100.0% | 100.0% | 83.0% | 43734 | 21409 | 100.0% | 49.0% |
| nan | 1 | 1 | 0 | 0.0% | 0.0% | 100.0% | 100.0% | 100.0% | 43735 | 21410 | 100.0% | 49.0% |



## 2.test

表5 等宽分箱数据分布表

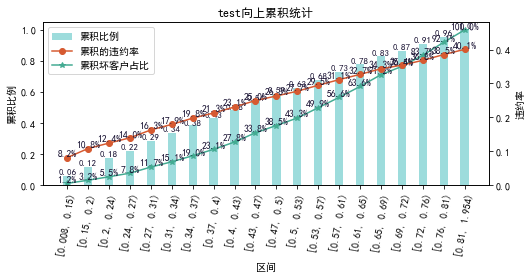
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [0.02, 0.1) | 865 | 62 | 803 | 0.5% | 4.2% | 0.5% | 4.2% | 7.2% | 865 | 62 | 2.7% | 7.2% |
| [0.1, 0.2) | 3213 | 387 | 2826 | 3.0% | 14.7% | 3.5% | 18.9% | 12.0% | 4078 | 449 | 12.7% | 11.0% |
| [0.2, 0.3) | 4582 | 912 | 3670 | 7.1% | 19.1% | 10.6% | 37.9% | 19.9% | 8660 | 1361 | 27.0% | 15.7% |
| [0.3, 0.4) | 4888 | 1464 | 3424 | 11.4% | 17.8% | 22.0% | 55.7% | 30.0% | 13548 | 2825 | 42.2% | 20.9% |
| [0.4, 0.5) | 4574 | 1836 | 2738 | 14.3% | 14.2% | 36.2% | 69.9% | 40.1% | 18122 | 4661 | 56.4% | 25.7% |
| [0.5, 0.6) | 3994 | 1941 | 2053 | 15.1% | 10.7% | 51.3% | 80.6% | 48.6% | 22116 | 6602 | 68.9% | 29.9% |
| [0.6, 0.7) | 3861 | 2135 | 1726 | 16.6% | 9.0% | 67.9% | 89.6% | 55.3% | 25977 | 8737 | 80.9% | 33.6% |
| [0.7, 0.8) | 3479 | 2186 | 1293 | 17.0% | 6.7% | 84.9% | 96.3% | 62.8% | 29456 | 10923 | 91.7% | 37.1% |
| [0.8, 0.9) | 2250 | 1628 | 622 | 12.7% | 3.2% | 97.6% | 99.5% | 72.4% | 31706 | 12551 | 98.7% | 39.6% |
| [0.9, 1.0) | 404 | 312 | 92 | 2.4% | 0.5% | 100.0% | 100.0% | 77.2% | 32110 | 12863 | 100.0% | 40.1% |

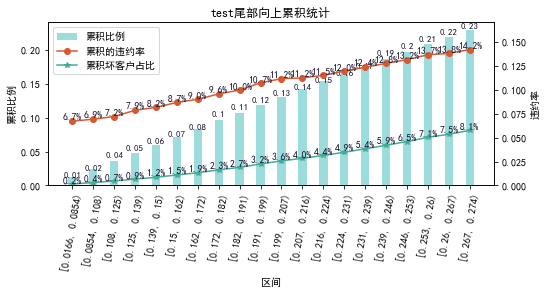


# 八、累积分数分布

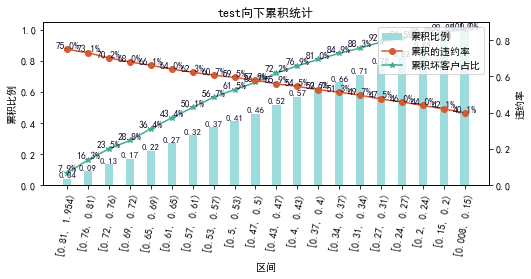
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

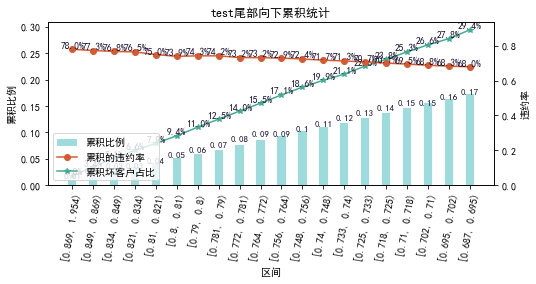
## 1. 向上累积分布





## 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 |
| [0.0167, 0.111) | 1.8% | 2.7% | 0.004 |
| [0.111, 0.205) | 7.4% | 10.0% | 0.008 |
| [0.205, 0.298) | 12.1% | 14.3% | 0.004 |
| [0.298, 0.392) | 14.4% | 15.2% | 0.001 |
| [0.392, 0.486) | 15.0% | 14.2% | 0.0 |
| [0.486, 0.579) | 12.9% | 12.4% | 0.0 |
| [0.579, 0.673) | 12.5% | 12.0% | 0.0 |
| [0.673, 0.766) | 13.2% | 10.8% | 0.004 |
| [0.766, 0.86) | 9.3% | 7.0% | 0.007 |
| [0.86, 0.954) | 1.5% | 1.3% | 0.0 |
| sum | 100.0% | 100.0% | 0.028 |

# 十、策略建议