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 |

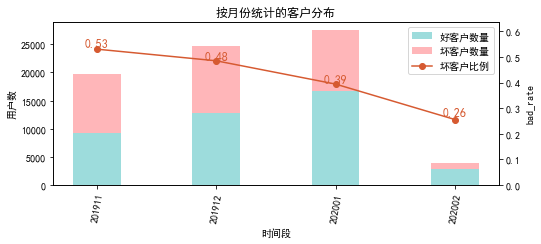
# 一、建模样本

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

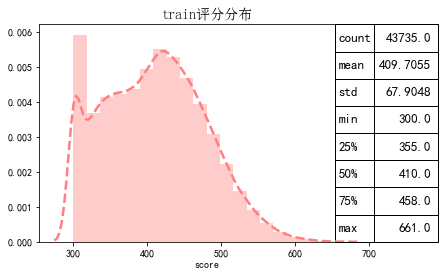
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 201911 | 9241 | 10434 | 19675 | 53.0% | 9241.0 | 10433.0 | 19674.0 | 53.0% | 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% |

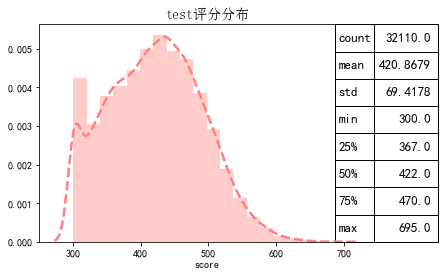


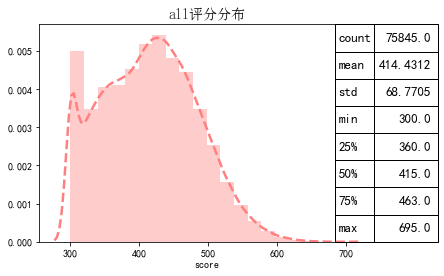
# 二、效果概况

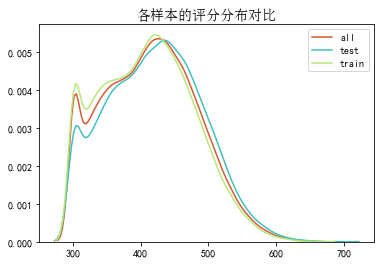
表3 模型效果概况表

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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 | 0.382 | 0.343 | 0.368 |
| AUC | 0.753 | 0.731 | 0.745 |



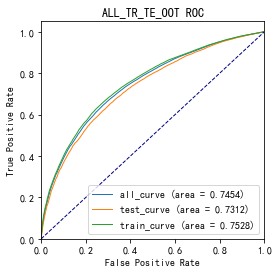
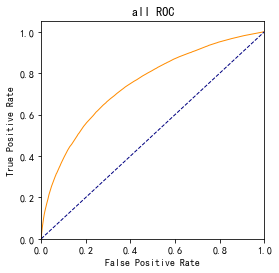
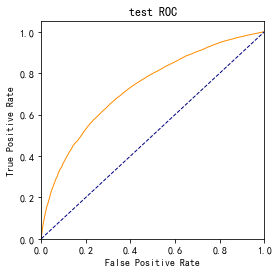
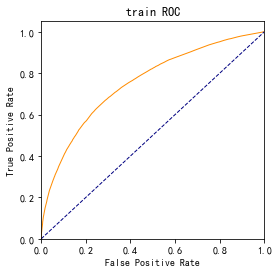






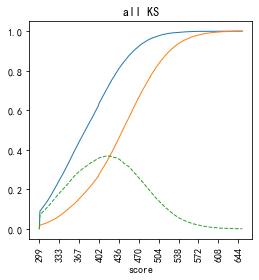
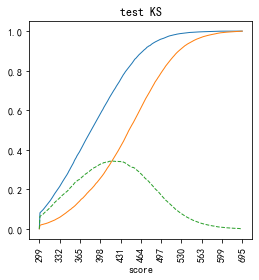
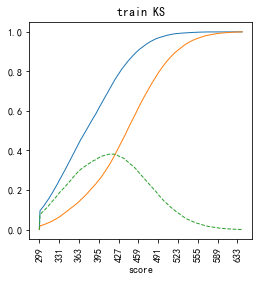
# 三、ROC曲线

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

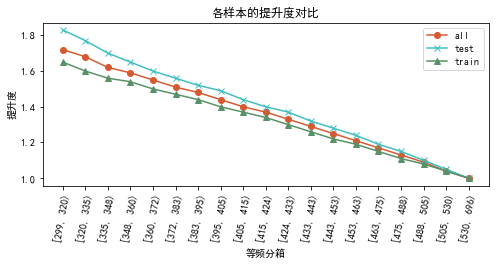
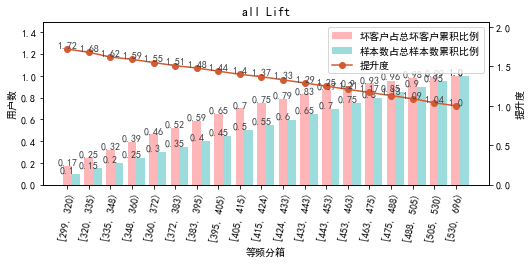
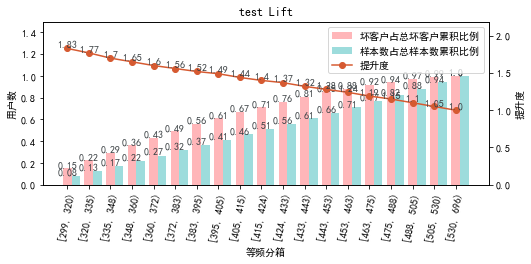
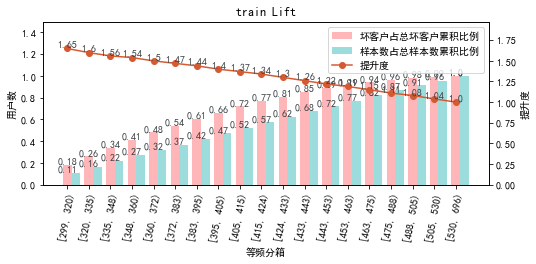
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

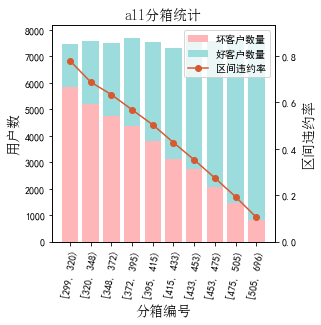
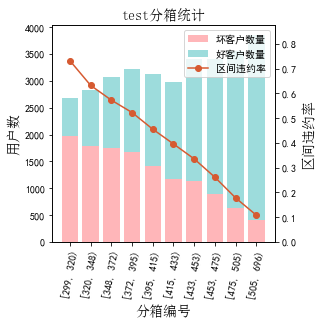
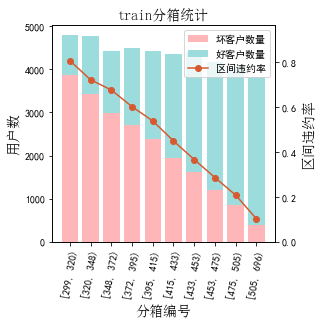
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



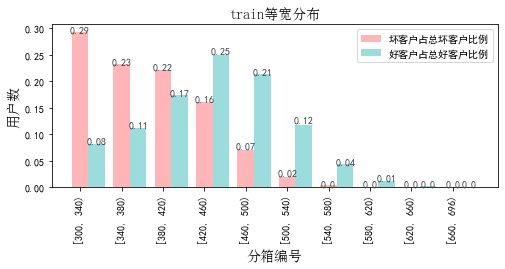
# 七、分数分布

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

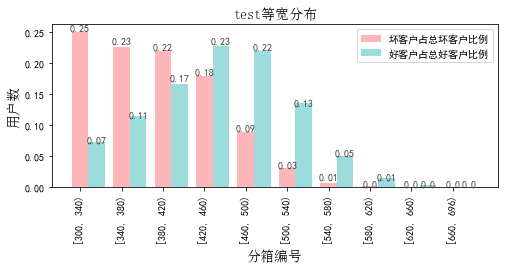
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [300, 340) | 8089 | 6272 | 1817 | 29.3% | 8.1% | 29.3% | 8.1% | 77.5% | 8089 | 6272 | 18.5% | 77.5% |
| [340, 380) | 7452 | 4978 | 2474 | 23.3% | 11.1% | 52.5% | 19.2% | 66.8% | 15541 | 11250 | 35.5% | 72.4% |
| [380, 420) | 8607 | 4736 | 3871 | 22.1% | 17.3% | 74.7% | 36.6% | 55.0% | 24148 | 15986 | 55.2% | 66.2% |
| [420, 460) | 8997 | 3426 | 5571 | 16.0% | 25.0% | 90.7% | 61.5% | 38.1% | 33145 | 19412 | 75.8% | 58.6% |
| [460, 500) | 6239 | 1497 | 4742 | 7.0% | 21.2% | 97.7% | 82.8% | 24.0% | 39384 | 20909 | 90.1% | 53.1% |
| [500, 540) | 3011 | 413 | 2598 | 1.9% | 11.6% | 99.6% | 94.4% | 13.7% | 42395 | 21322 | 96.9% | 50.3% |
| [540, 580) | 1043 | 83 | 960 | 0.4% | 4.3% | 100.0% | 98.7% | 8.0% | 43438 | 21405 | 99.3% | 49.3% |
| [580, 620) | 256 | 4 | 252 | 0.0% | 1.1% | 100.0% | 99.8% | 1.6% | 43694 | 21409 | 99.9% | 49.0% |
| [620, 660) | 39 | 1 | 38 | 0.0% | 0.2% | 100.0% | 100.0% | 2.6% | 43733 | 21410 | 100.0% | 49.0% |
| [660, 696) | 2 | 0 | 2 | 0.0% | 0.0% | 100.0% | 100.0% | 0.0% | 43735 | 21410 | 100.0% | 49.0% |



## 2.test

表5 等宽分箱数据分布表

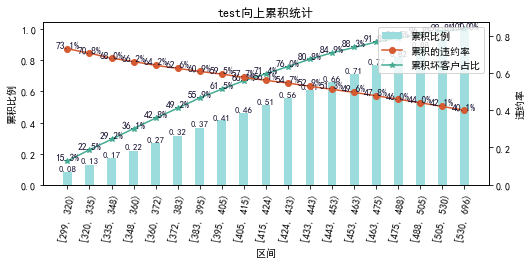
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [300, 340) | 4618 | 3215 | 1403 | 25.0% | 7.3% | 25.0% | 7.3% | 69.6% | 4618 | 3215 | 14.4% | 69.6% |
| [340, 380) | 5104 | 2898 | 2206 | 22.5% | 11.5% | 47.5% | 18.8% | 56.8% | 9722 | 6113 | 30.3% | 62.9% |
| [380, 420) | 5999 | 2822 | 3177 | 21.9% | 16.5% | 69.5% | 35.3% | 47.0% | 15721 | 8935 | 49.0% | 56.8% |
| [420, 460) | 6679 | 2300 | 4379 | 17.9% | 22.8% | 87.3% | 58.0% | 34.4% | 22400 | 11235 | 69.8% | 50.2% |
| [460, 500) | 5336 | 1133 | 4203 | 8.8% | 21.8% | 96.2% | 79.8% | 21.2% | 27736 | 12368 | 86.4% | 44.6% |
| [500, 540) | 2981 | 391 | 2590 | 3.0% | 13.5% | 99.2% | 93.3% | 13.1% | 30717 | 12759 | 95.7% | 41.5% |
| [540, 580) | 1039 | 80 | 959 | 0.6% | 5.0% | 99.8% | 98.3% | 7.7% | 31756 | 12839 | 98.9% | 40.4% |
| [580, 620) | 288 | 21 | 267 | 0.2% | 1.4% | 100.0% | 99.7% | 7.3% | 32044 | 12860 | 99.8% | 40.1% |
| [620, 660) | 60 | 2 | 58 | 0.0% | 0.3% | 100.0% | 100.0% | 3.3% | 32104 | 12862 | 100.0% | 40.1% |
| [660, 696) | 6 | 1 | 5 | 0.0% | 0.0% | 100.0% | 100.0% | 16.7% | 32110 | 12863 | 100.0% | 40.1% |

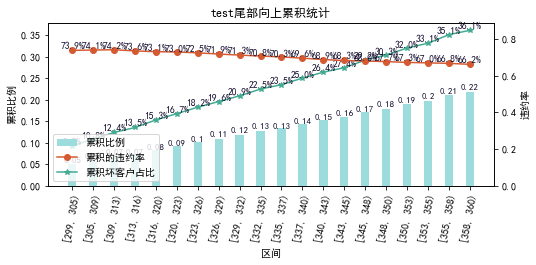


# 八、累积分数分布

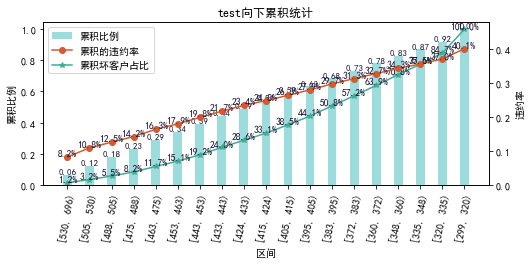
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

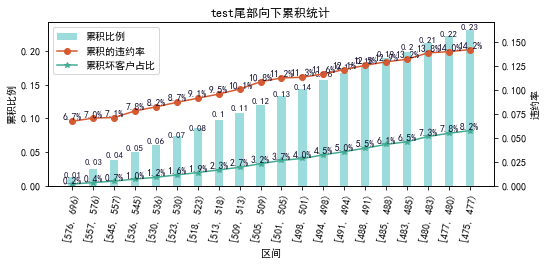
## 1. 向上累积分布





## 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 |
| [299.605, 339.5) | 18.5% | 14.4% | 0.01 |
| [339.5, 379.0) | 16.6% | 15.4% | 0.001 |
| [379.0, 418.5) | 19.6% | 18.6% | 0.0 |
| [418.5, 458.0) | 20.2% | 20.3% | 0.0 |
| [458.0, 497.5) | 14.7% | 17.0% | 0.003 |
| [497.5, 537.0) | 7.1% | 9.5% | 0.007 |
| [537.0, 576.5) | 2.6% | 3.5% | 0.003 |
| [576.5, 616.0) | 0.6% | 1.0% | 0.002 |
| [616.0, 655.5) | 0.1% | 0.2% | 0.001 |
| [655.5, 695.0) | 0.0% | 0.0% | 0.0 |
| sum | 100.0% | 100.0% | 0.027 |

# 十、策略建议