AKULAKU：id\_pdlnew\_17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 | 2020-05-09 |

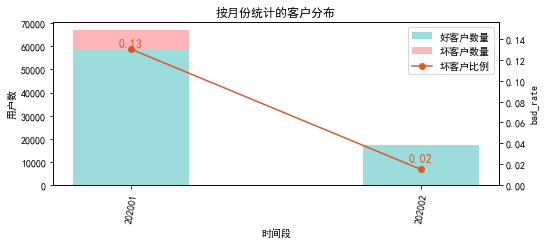
# 一、建模样本

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 |
|  | 84320 | 0.1068 | 202001 | 202002 | -- | -- |

表2 按月份统计分布表

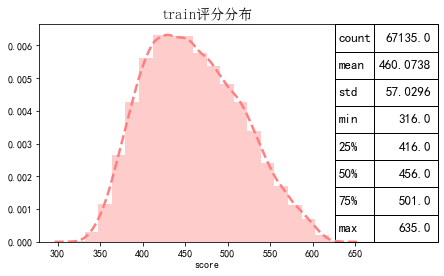
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 202001 | 58386.0 | 8749.0 | 67135 | 13.0% | 58386.0 | 8749.0 | 67135.0 | 13.0% | 0.0 | 0.0 | 0.0 | 0.0% |
| 202002 | 16926.0 | 259.0 | 17185 | 1.5% | 0.0 | 0.0 | 0.0 | 0.0% | 16926.0 | 259.0 | 17185.0 | 1.5% |

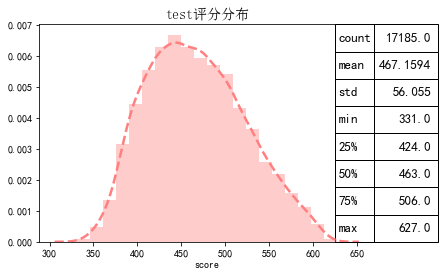


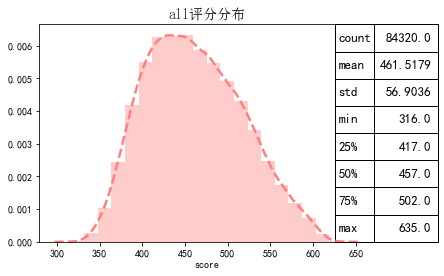
# 二、效果概况

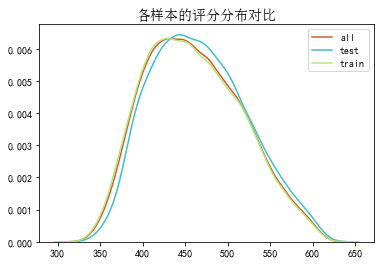
表3 模型效果概况表

|  |  |  |  |
| --- | --- | --- | --- |
|  | train | test | ALL |
| 样本量 | 67135 | 17185 | 84320 |
| 坏客户数量 | 8749 | 259 | 9008 |
| 坏客户比例 | 13.0% | 1.5% | 10.7% |
| KS | 0.356 | 0.349 | 0.357 |
| AUC | 0.74 | 0.725 | 0.742 |



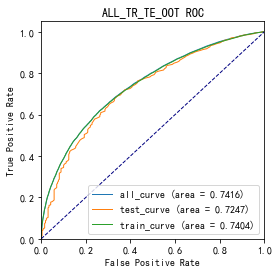
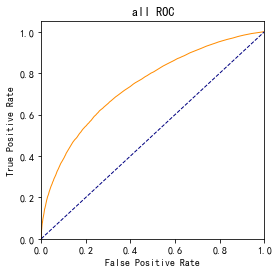
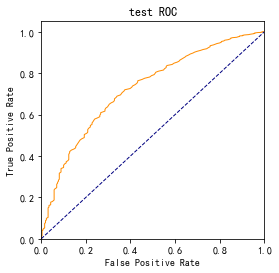
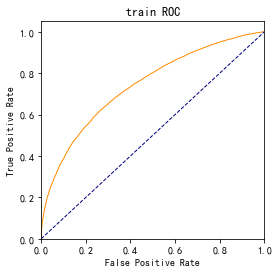






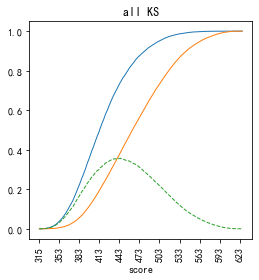
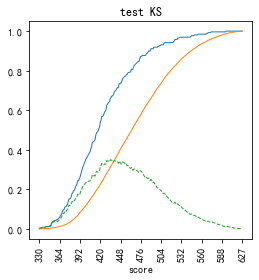
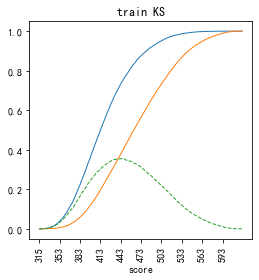
# 三、ROC曲线

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

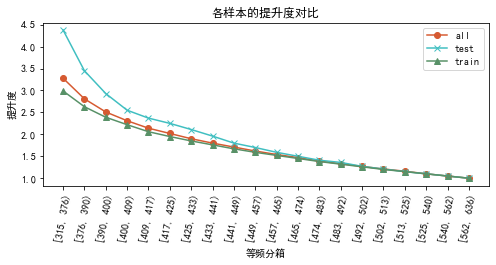
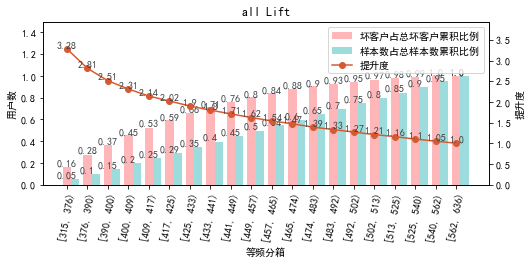
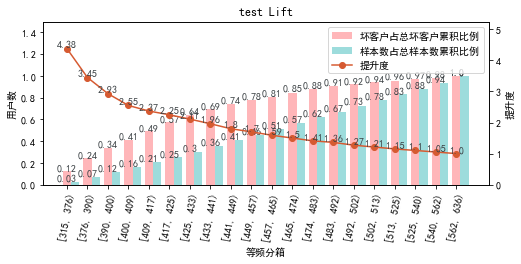
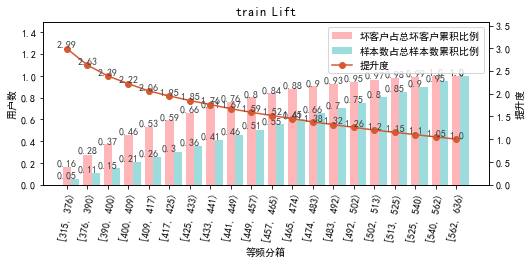
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

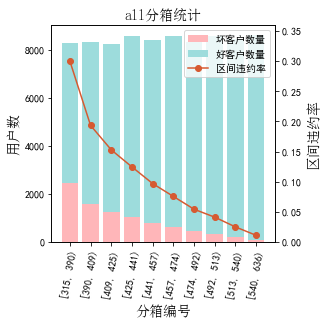
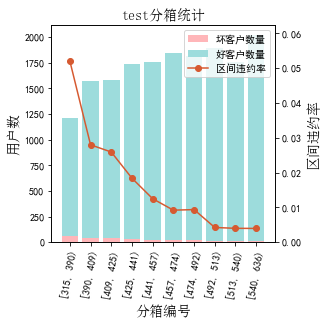
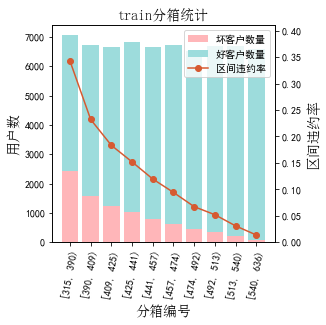
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



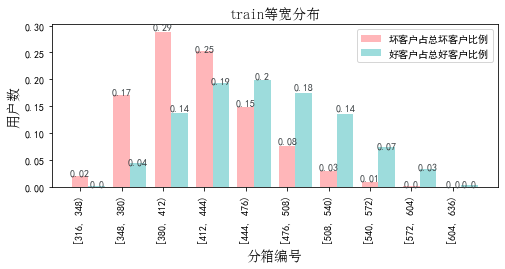
# 七、分数分布

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

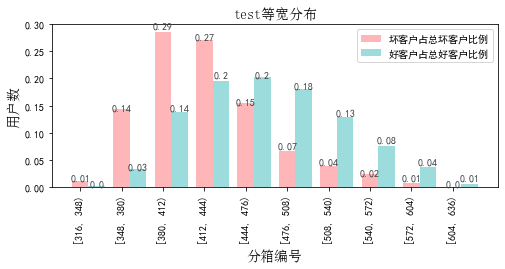
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [316, 348) | 343 | 177 | 166 | 2.0% | 0.3% | 2.0% | 0.3% | 51.6% | 343 | 177 | 0.5% | 51.6% |
| [348, 380) | 4076 | 1494 | 2582 | 17.1% | 4.4% | 19.1% | 4.7% | 36.7% | 4419 | 1671 | 6.6% | 37.8% |
| [380, 412) | 10577 | 2524 | 8053 | 28.8% | 13.8% | 47.9% | 18.5% | 23.9% | 14996 | 4195 | 22.3% | 28.0% |
| [412, 444) | 13514 | 2219 | 11295 | 25.4% | 19.3% | 73.3% | 37.8% | 16.4% | 28510 | 6414 | 42.5% | 22.5% |
| [444, 476) | 12919 | 1306 | 11613 | 14.9% | 19.9% | 88.2% | 57.7% | 10.1% | 41429 | 7720 | 61.7% | 18.6% |
| [476, 508) | 10899 | 670 | 10229 | 7.7% | 17.5% | 95.9% | 75.3% | 6.1% | 52328 | 8390 | 77.9% | 16.0% |
| [508, 540) | 8216 | 269 | 7947 | 3.1% | 13.6% | 99.0% | 88.9% | 3.3% | 60544 | 8659 | 90.2% | 14.3% |
| [540, 572) | 4389 | 78 | 4311 | 0.9% | 7.4% | 99.9% | 96.2% | 1.8% | 64933 | 8737 | 96.7% | 13.5% |
| [572, 604) | 1958 | 11 | 1947 | 0.1% | 3.3% | 100.0% | 99.6% | 0.6% | 66891 | 8748 | 99.6% | 13.1% |
| [604, 636) | 244 | 1 | 243 | 0.0% | 0.4% | 100.0% | 100.0% | 0.4% | 67135 | 8749 | 100.0% | 13.0% |



## 2.test

表5 等宽分箱数据分布表

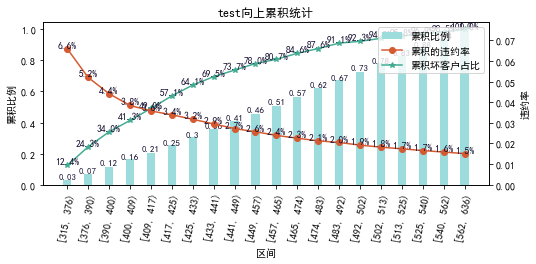
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [316, 348) | 33 | 3 | 30 | 1.2% | 0.2% | 1.2% | 0.2% | 9.1% | 33 | 3 | 0.2% | 9.1% |
| [348, 380) | 605 | 37 | 568 | 14.3% | 3.4% | 15.4% | 3.5% | 6.1% | 638 | 40 | 3.7% | 6.3% |
| [380, 412) | 2406 | 74 | 2332 | 28.6% | 13.8% | 44.0% | 17.3% | 3.1% | 3044 | 114 | 17.7% | 3.7% |
| [412, 444) | 3381 | 70 | 3311 | 27.0% | 19.6% | 71.0% | 36.9% | 2.1% | 6425 | 184 | 37.4% | 2.9% |
| [444, 476) | 3475 | 40 | 3435 | 15.4% | 20.3% | 86.5% | 57.2% | 1.2% | 9900 | 224 | 57.6% | 2.3% |
| [476, 508) | 3084 | 17 | 3067 | 6.6% | 18.1% | 93.1% | 75.3% | 0.6% | 12984 | 241 | 75.6% | 1.9% |
| [508, 540) | 2182 | 10 | 2172 | 3.9% | 12.8% | 96.9% | 88.1% | 0.5% | 15166 | 251 | 88.3% | 1.7% |
| [540, 572) | 1294 | 6 | 1288 | 2.3% | 7.6% | 99.2% | 95.7% | 0.5% | 16460 | 257 | 95.8% | 1.6% |
| [572, 604) | 629 | 2 | 627 | 0.8% | 3.7% | 100.0% | 99.4% | 0.3% | 17089 | 259 | 99.4% | 1.5% |
| [604, 636) | 96 | 0 | 96 | 0.0% | 0.6% | 100.0% | 100.0% | 0.0% | 17185 | 259 | 100.0% | 1.5% |

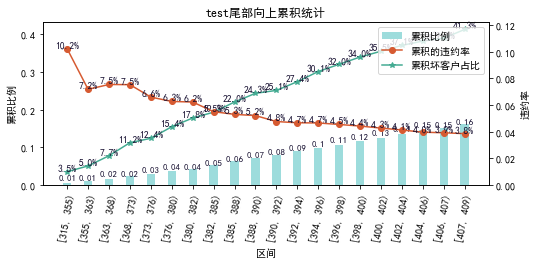


# 八、累积分数分布

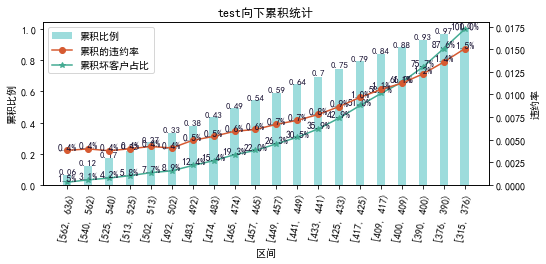
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

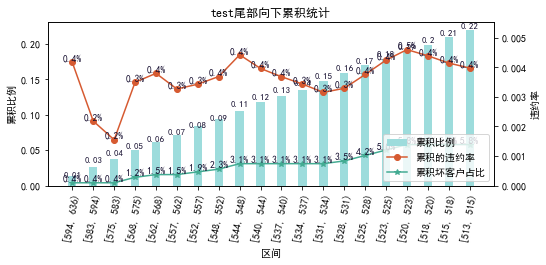
## 1. 向上累积分布





## 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 |
| [315.681, 347.9) | 0.5% | 0.2% | 0.003 |
| [347.9, 379.8) | 6.1% | 3.5% | 0.014 |
| [379.8, 411.7) | 15.8% | 14.0% | 0.002 |
| [411.7, 443.6) | 20.1% | 19.7% | 0.0 |
| [443.6, 475.5) | 19.2% | 20.2% | 0.0 |
| [475.5, 507.4) | 16.2% | 17.9% | 0.002 |
| [507.4, 539.3) | 12.2% | 12.7% | 0.0 |
| [539.3, 571.2) | 6.5% | 7.5% | 0.001 |
| [571.2, 603.1) | 2.9% | 3.7% | 0.002 |
| [603.1, 635.0) | 0.4% | 0.6% | 0.001 |
| sum | 100.0% | 100.0% | 0.026 |

# 十、策略建议