AKULAKU：idpdlold87(改)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 | 2020-10-14 |

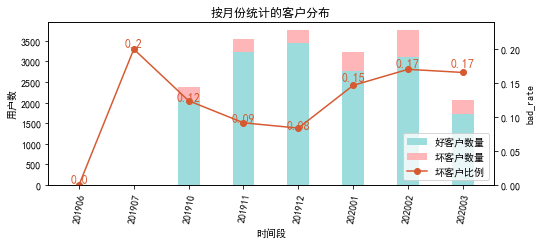
# 一、建模样本

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 |
|  | 18750 | 0.1276 | 201906 | 202002 | 202003 | 202003 |

表2 按月份统计分布表

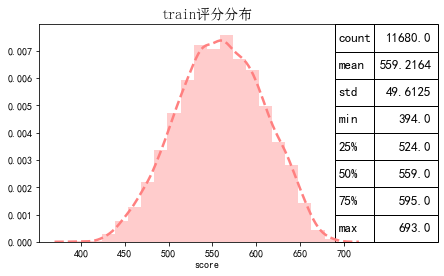
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 201906 | 5 | 0 | 5 | 0.0% | 3.0 | 0.0 | 3.0 | 0.0% | 2.0 | 0.0 | 2.0 | 0.0% |
| 201907 | 4 | 1 | 5 | 20.0% | 3.0 | 1.0 | 4.0 | 25.0% | 1.0 | 0.0 | 1.0 | 0.0% |
| 201910 | 2076 | 294 | 2370 | 12.4% | 1483.0 | 210.0 | 1693.0 | 12.4% | 593.0 | 84.0 | 677.0 | 12.4% |
| 201911 | 3223 | 325 | 3548 | 9.2% | 2217.0 | 233.0 | 2450.0 | 9.5% | 1006.0 | 92.0 | 1098.0 | 8.4% |
| 201912 | 3453 | 316 | 3769 | 8.4% | 2441.0 | 205.0 | 2646.0 | 7.7% | 1012.0 | 111.0 | 1123.0 | 9.9% |
| 202001 | 2757 | 474 | 3231 | 14.7% | 1908.0 | 340.0 | 2248.0 | 15.1% | 849.0 | 134.0 | 983.0 | 13.6% |
| 202002 | 3118 | 640 | 3758 | 17.0% | 2190.0 | 446.0 | 2636.0 | 16.9% | 928.0 | 194.0 | 1122.0 | 17.3% |
| 202003 | 1722 | 342 | 2064 | 16.6% | 0.0 | 0.0 | 0.0 | 0.0% | 0.0 | 0.0 | 0.0 | 0.0% |

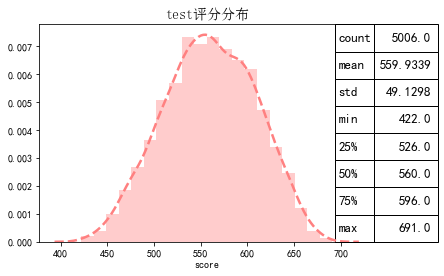


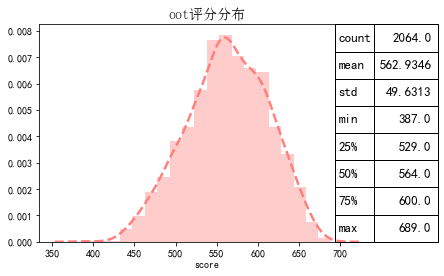
# 二、效果概况

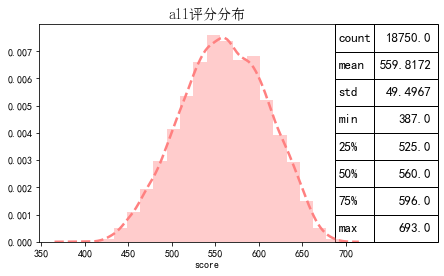
表3 模型效果概况表

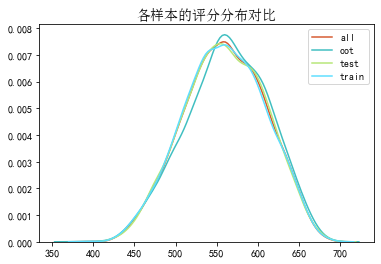
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | train | test | oot | ALL |
| 样本量 | 11680 | 5006 | 2064 | 18750 |
| 坏客户数量 | 1435 | 615 | 342 | 2392 |
| 坏客户比例 | 12.3% | 12.3% | 16.6% | 12.8% |
| KS | 0.325 | 0.299 | 0.243 | 0.3 |
| AUC | 0.714 | 0.691 | 0.656 | 0.699 |





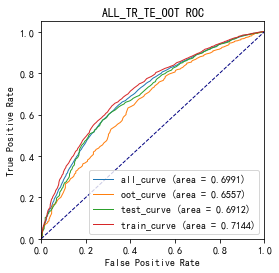
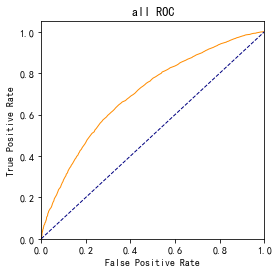
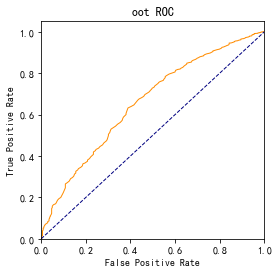
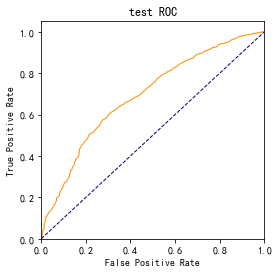
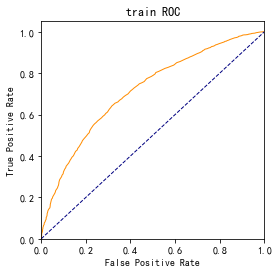






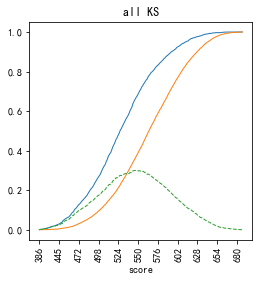
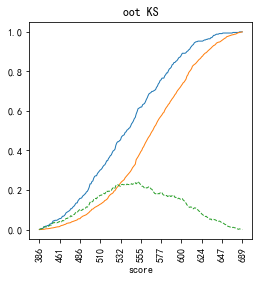
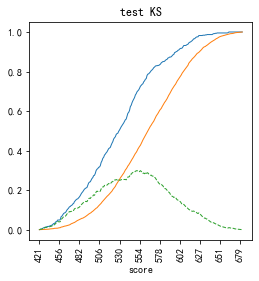
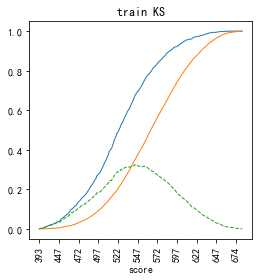
# 三、ROC曲线

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

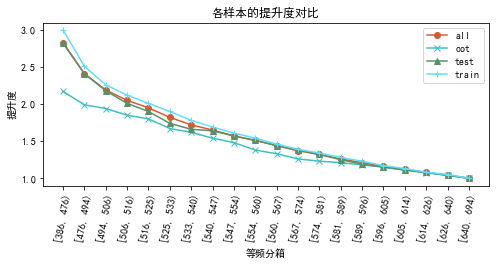
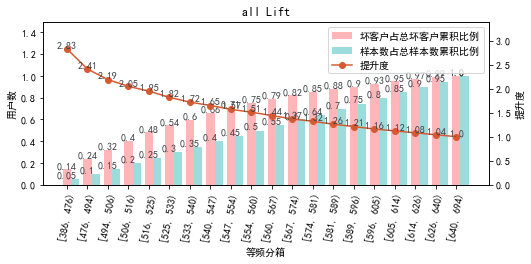
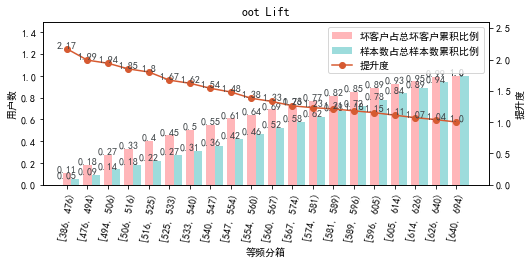
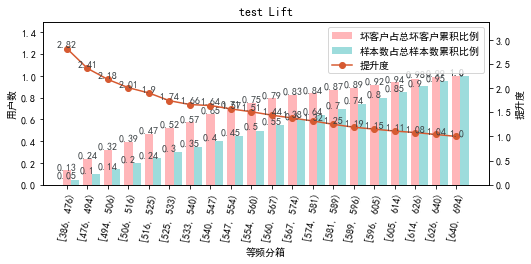
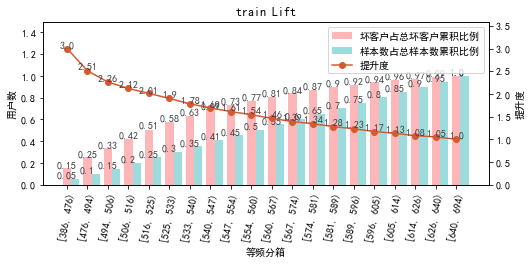
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

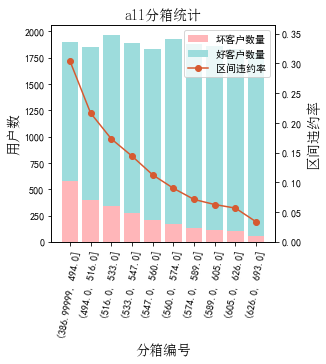
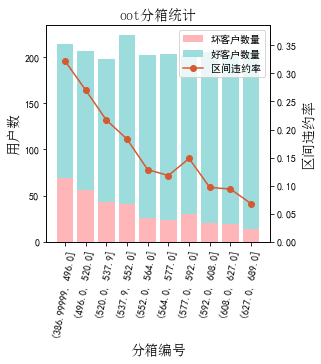
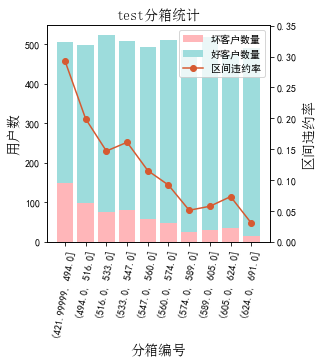
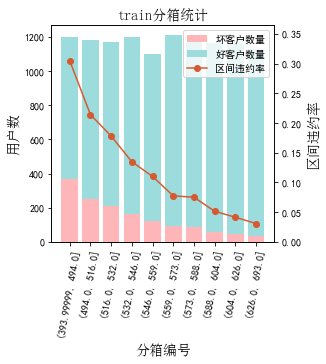
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



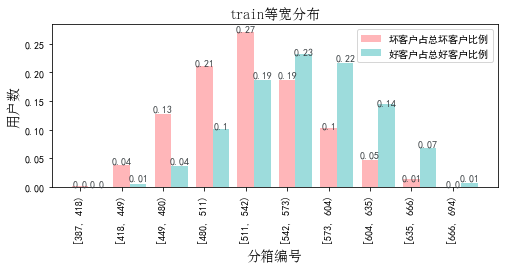
# 七、分数分布

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

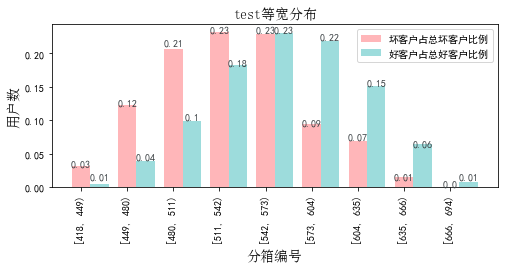
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [387, 418) | 4 | 3 | 1 | 0.2% | 0.0% | 0.2% | 0.0% | 75.0% | 4 | 3 | 0.0% | 75.0% |
| [418, 449) | 114 | 55 | 59 | 3.8% | 0.6% | 4.0% | 0.6% | 48.2% | 118 | 58 | 1.0% | 49.2% |
| [449, 480) | 564 | 182 | 382 | 12.7% | 3.7% | 16.7% | 4.3% | 32.3% | 682 | 240 | 5.8% | 35.2% |
| [480, 511) | 1335 | 302 | 1033 | 21.0% | 10.1% | 37.8% | 14.4% | 22.6% | 2017 | 542 | 17.3% | 26.9% |
| [511, 542) | 2302 | 389 | 1913 | 27.1% | 18.7% | 64.9% | 33.1% | 16.9% | 4319 | 931 | 37.0% | 21.6% |
| [542, 573) | 2652 | 269 | 2383 | 18.7% | 23.3% | 83.6% | 56.3% | 10.1% | 6971 | 1200 | 59.7% | 17.2% |
| [573, 604) | 2360 | 147 | 2213 | 10.2% | 21.6% | 93.9% | 77.9% | 6.2% | 9331 | 1347 | 79.9% | 14.4% |
| [604, 635) | 1550 | 68 | 1482 | 4.7% | 14.5% | 98.6% | 92.4% | 4.4% | 10881 | 1415 | 93.2% | 13.0% |
| [635, 666) | 723 | 20 | 703 | 1.4% | 6.9% | 100.0% | 99.3% | 2.8% | 11604 | 1435 | 99.3% | 12.4% |
| [666, 694) | 76 | 0 | 76 | 0.0% | 0.7% | 100.0% | 100.0% | 0.0% | 11680 | 1435 | 100.0% | 12.3% |



## 2.test

表5 等宽分箱数据分布表

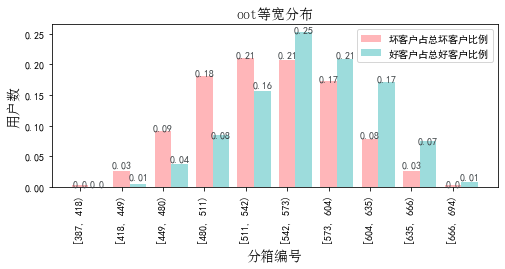
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [418, 449) | 42 | 19 | 23 | 3.1% | 0.5% | 3.1% | 0.5% | 45.2% | 42 | 19 | 0.8% | 45.2% |
| [449, 480) | 250 | 76 | 174 | 12.4% | 4.0% | 15.4% | 4.5% | 30.4% | 292 | 95 | 5.8% | 32.5% |
| [480, 511) | 563 | 127 | 436 | 20.7% | 9.9% | 36.1% | 14.4% | 22.6% | 855 | 222 | 17.1% | 26.0% |
| [511, 542) | 947 | 143 | 804 | 23.3% | 18.3% | 59.3% | 32.7% | 15.1% | 1802 | 365 | 36.0% | 20.3% |
| [542, 573) | 1154 | 141 | 1013 | 22.9% | 23.1% | 82.3% | 55.8% | 12.2% | 2956 | 506 | 59.0% | 17.1% |
| [573, 604) | 1016 | 58 | 958 | 9.4% | 21.8% | 91.7% | 77.6% | 5.7% | 3972 | 564 | 79.3% | 14.2% |
| [604, 635) | 706 | 42 | 664 | 6.8% | 15.1% | 98.5% | 92.7% | 5.9% | 4678 | 606 | 93.4% | 13.0% |
| [635, 666) | 292 | 9 | 283 | 1.5% | 6.4% | 100.0% | 99.2% | 3.1% | 4970 | 615 | 99.3% | 12.4% |
| [666, 694) | 36 | 0 | 36 | 0.0% | 0.8% | 100.0% | 100.0% | 0.0% | 5006 | 615 | 100.0% | 12.3% |



## 3.oot

表6 等宽分箱数据分布表

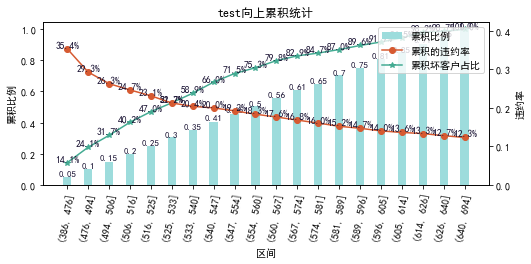
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [387, 418) | 2 | 1 | 1 | 0.3% | 0.1% | 0.3% | 0.1% | 50.0% | 2 | 1 | 0.1% | 50.0% |
| [418, 449) | 19 | 9 | 10 | 2.6% | 0.6% | 2.9% | 0.6% | 47.4% | 21 | 10 | 1.0% | 47.6% |
| [449, 480) | 95 | 31 | 64 | 9.1% | 3.7% | 12.0% | 4.4% | 32.6% | 116 | 41 | 5.6% | 35.3% |
| [480, 511) | 207 | 62 | 145 | 18.1% | 8.4% | 30.1% | 12.8% | 30.0% | 323 | 103 | 15.6% | 31.9% |
| [511, 542) | 341 | 72 | 269 | 21.1% | 15.6% | 51.2% | 28.4% | 21.1% | 664 | 175 | 32.2% | 26.4% |
| [542, 573) | 507 | 71 | 436 | 20.8% | 25.3% | 71.9% | 53.7% | 14.0% | 1171 | 246 | 56.7% | 21.0% |
| [573, 604) | 419 | 59 | 360 | 17.3% | 20.9% | 89.2% | 74.6% | 14.1% | 1590 | 305 | 77.0% | 19.2% |
| [604, 635) | 321 | 27 | 294 | 7.9% | 17.1% | 97.1% | 91.7% | 8.4% | 1911 | 332 | 92.6% | 17.4% |
| [635, 666) | 138 | 9 | 129 | 2.6% | 7.5% | 99.7% | 99.2% | 6.5% | 2049 | 341 | 99.3% | 16.6% |
| [666, 694) | 15 | 1 | 14 | 0.3% | 0.8% | 100.0% | 100.0% | 6.7% | 2064 | 342 | 100.0% | 16.6% |

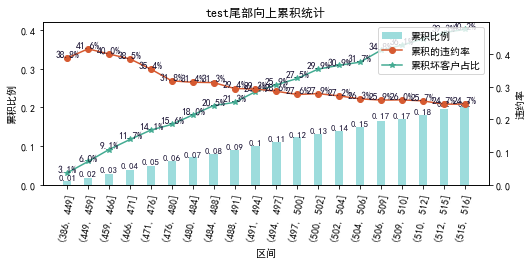


# 八、累积分数分布

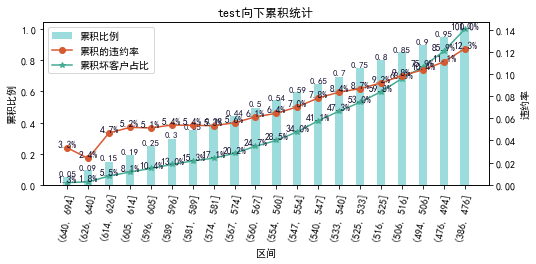
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

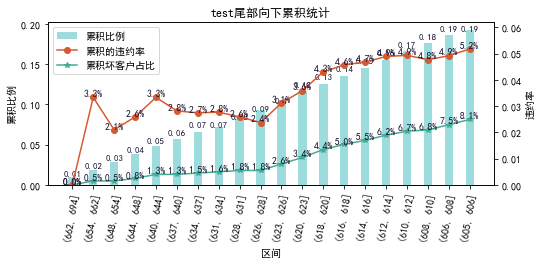
## 1. 向上累积分布





## 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 | oot\_pct | train\_test\_psi | train\_oot\_psi |
| [386.694, 417.6) | 0.0% | 0.0% | 0.1% |  | 0.001 |
| [417.6, 448.2) | 1.0% | 0.8% | 0.9% | 0.0 | 0.0 |
| [448.2, 478.8) | 4.6% | 4.7% | 4.4% | 0.0 | 0.0 |
| [478.8, 509.4) | 11.0% | 11.0% | 9.9% | 0.0 | 0.001 |
| [509.4, 540.0) | 18.8% | 18.1% | 15.7% | 0.0 | 0.005 |
| [540.0, 570.6) | 22.7% | 23.0% | 24.1% | 0.0 | 0.001 |
| [570.6, 601.2) | 20.6% | 20.6% | 20.9% | 0.0 | 0.0 |
| [601.2, 631.8) | 13.5% | 14.7% | 15.4% | 0.001 | 0.003 |
| [631.8, 662.4) | 6.8% | 6.3% | 7.4% | 0.0 | 0.0 |
| [662.4, 693.0) | 0.9% | 0.9% | 1.1% | 0.0 | 0.0 |
| sum | 100.0% | 100.0% | 100.0% | 0.002 | 0.012 |

# 十、策略建议