AKULAKU：idpdlold87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 | 2020-10-13 |

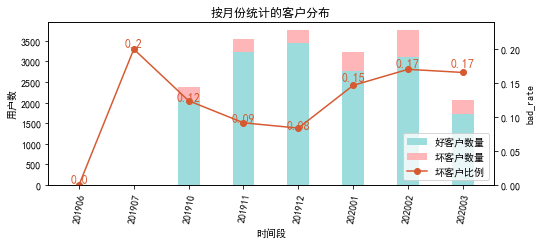
# 一、建模样本

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

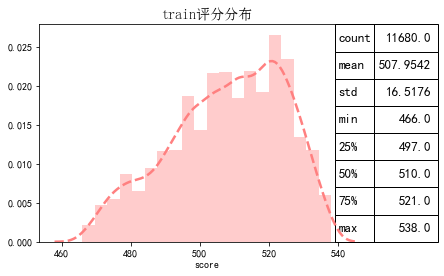
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 201906 | 5 | 0 | 5 | 0.0% | 4.0 | 0.0 | 4.0 | 0.0% | 1.0 | 0.0 | 1.0 | 0.0% |
| 201907 | 4 | 1 | 5 | 20.0% | 1.0 | 1.0 | 2.0 | 50.0% | 3.0 | 0.0 | 3.0 | 0.0% |
| 201910 | 2076 | 294 | 2370 | 12.4% | 1445.0 | 210.0 | 1655.0 | 12.7% | 631.0 | 84.0 | 715.0 | 11.7% |
| 201911 | 3223 | 325 | 3548 | 9.2% | 2258.0 | 247.0 | 2505.0 | 9.9% | 965.0 | 78.0 | 1043.0 | 7.5% |
| 201912 | 3453 | 316 | 3769 | 8.4% | 2420.0 | 217.0 | 2637.0 | 8.2% | 1033.0 | 99.0 | 1132.0 | 8.7% |
| 202001 | 2757 | 474 | 3231 | 14.7% | 1935.0 | 326.0 | 2261.0 | 14.4% | 822.0 | 148.0 | 970.0 | 15.3% |
| 202002 | 3118 | 640 | 3758 | 17.0% | 2182.0 | 434.0 | 2616.0 | 16.6% | 936.0 | 206.0 | 1142.0 | 18.0% |
| 202003 | 1722 | 342 | 2064 | 16.6% | 0.0 | 0.0 | 0.0 | 0.0% | 0.0 | 0.0 | 0.0 | 0.0% |

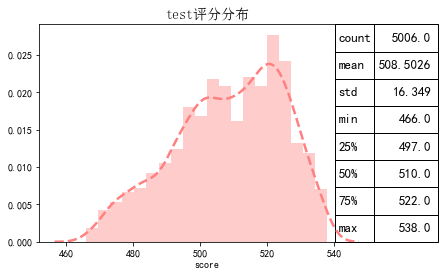


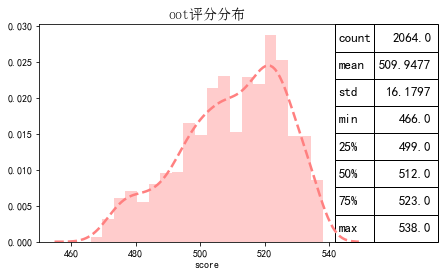
# 二、效果概况

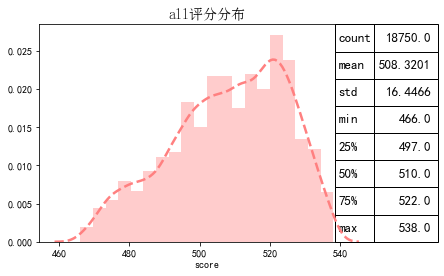
表3 模型效果概况表

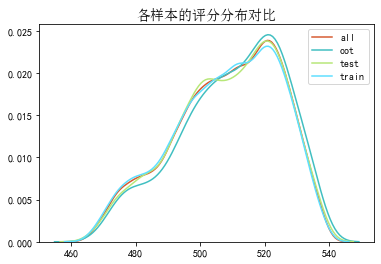
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | train | test | oot | ALL |
| 样本量 | 11680 | 5006 | 2064 | 18750 |
| 坏客户数量 | 1435 | 615 | 342 | 2392 |
| 坏客户比例 | 12.3% | 12.3% | 16.6% | 12.8% |
| KS | 0.279 | 0.27 | 0.23 | 0.263 |
| AUC | 0.687 | 0.677 | 0.64 | 0.676 |





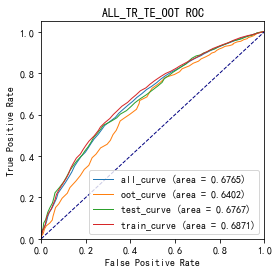
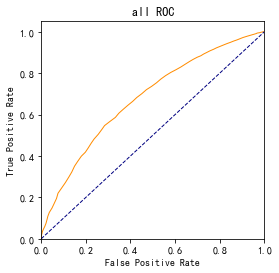
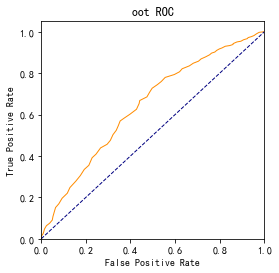
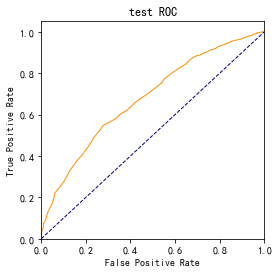
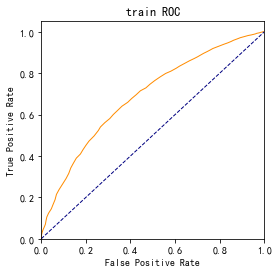






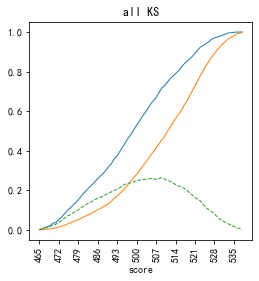
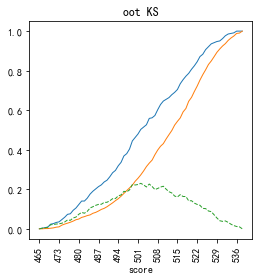
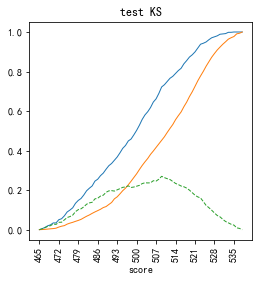
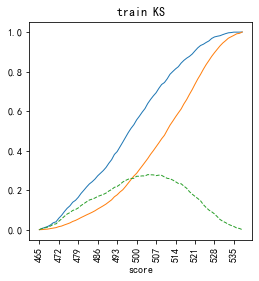
# 三、ROC曲线

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

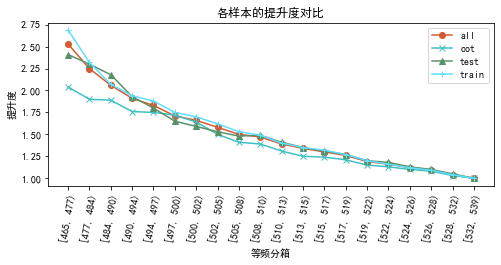
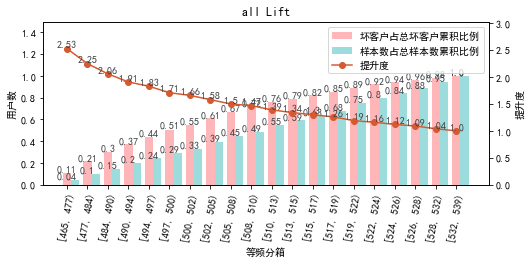
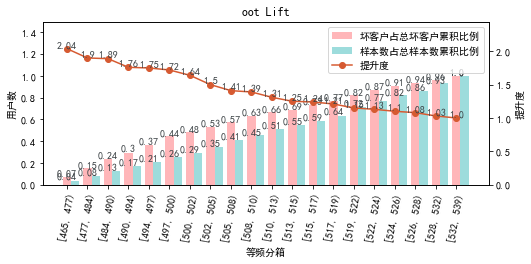
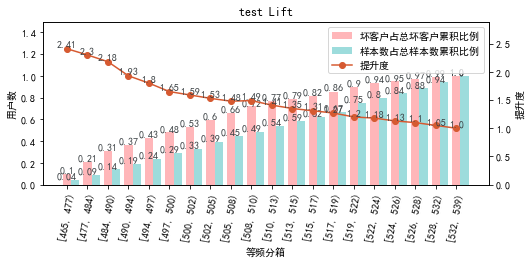
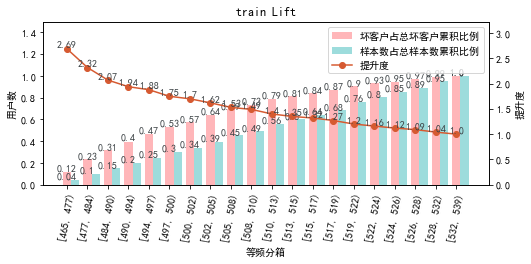
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

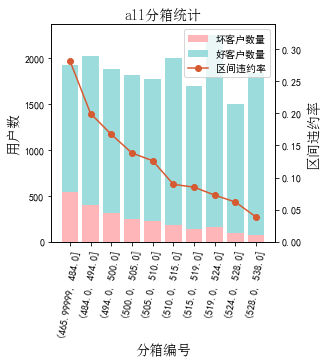
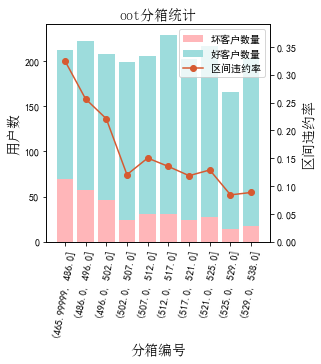
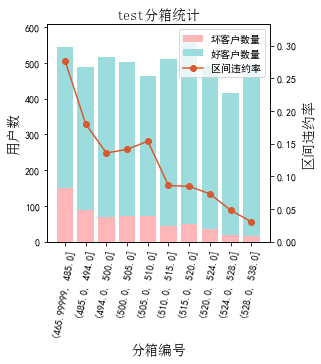
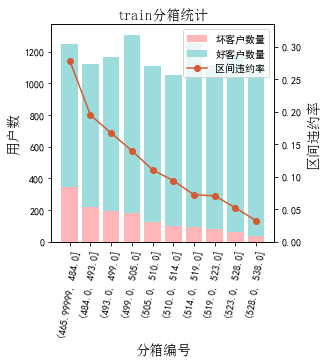
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



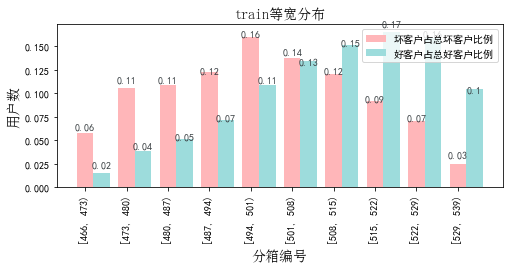
# 七、分数分布

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

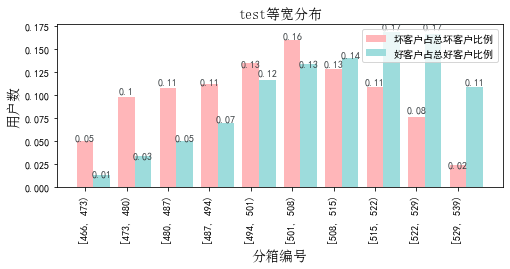
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [466, 473) | 239 | 83 | 156 | 5.8% | 1.5% | 5.8% | 1.5% | 34.7% | 239 | 83 | 2.0% | 34.7% |
| [473, 480) | 546 | 152 | 394 | 10.6% | 3.8% | 16.4% | 5.4% | 27.8% | 785 | 235 | 6.7% | 29.9% |
| [480, 487) | 676 | 156 | 520 | 10.9% | 5.1% | 27.2% | 10.4% | 23.1% | 1461 | 391 | 12.5% | 26.8% |
| [487, 494) | 913 | 176 | 737 | 12.3% | 7.2% | 39.5% | 17.6% | 19.3% | 2374 | 567 | 20.3% | 23.9% |
| [494, 501) | 1344 | 230 | 1114 | 16.0% | 10.9% | 55.5% | 28.5% | 17.1% | 3718 | 797 | 31.8% | 21.4% |
| [501, 508) | 1572 | 197 | 1375 | 13.7% | 13.4% | 69.3% | 41.9% | 12.5% | 5290 | 994 | 45.3% | 18.8% |
| [508, 515) | 1722 | 173 | 1549 | 12.1% | 15.1% | 81.3% | 57.1% | 10.0% | 7012 | 1167 | 60.0% | 16.6% |
| [515, 522) | 1828 | 131 | 1697 | 9.1% | 16.6% | 90.5% | 73.6% | 7.2% | 8840 | 1298 | 75.7% | 14.7% |
| [522, 529) | 1730 | 101 | 1629 | 7.0% | 15.9% | 97.5% | 89.5% | 5.8% | 10570 | 1399 | 90.5% | 13.2% |
| [529, 539) | 1110 | 36 | 1074 | 2.5% | 10.5% | 100.0% | 100.0% | 3.2% | 11680 | 1435 | 100.0% | 12.3% |



## 2.test

表5 等宽分箱数据分布表

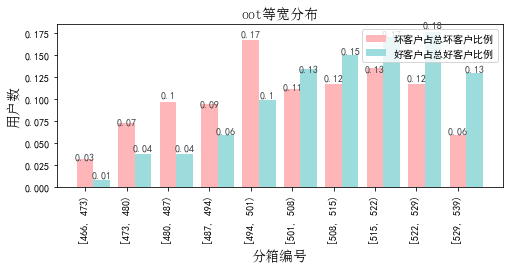
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [466, 473) | 89 | 31 | 58 | 5.0% | 1.3% | 5.0% | 1.3% | 34.8% | 89 | 31 | 1.8% | 34.8% |
| [473, 480) | 207 | 60 | 147 | 9.8% | 3.3% | 14.8% | 4.7% | 29.0% | 296 | 91 | 5.9% | 30.7% |
| [480, 487) | 284 | 66 | 218 | 10.7% | 5.0% | 25.5% | 9.6% | 23.2% | 580 | 157 | 11.6% | 27.1% |
| [487, 494) | 375 | 69 | 306 | 11.2% | 7.0% | 36.7% | 16.6% | 18.4% | 955 | 226 | 19.1% | 23.7% |
| [494, 501) | 595 | 83 | 512 | 13.5% | 11.7% | 50.2% | 28.3% | 13.9% | 1550 | 309 | 31.0% | 19.9% |
| [501, 508) | 684 | 98 | 586 | 15.9% | 13.3% | 66.2% | 41.6% | 14.3% | 2234 | 407 | 44.6% | 18.2% |
| [508, 515) | 695 | 79 | 616 | 12.8% | 14.0% | 79.0% | 55.6% | 11.4% | 2929 | 486 | 58.5% | 16.6% |
| [515, 522) | 807 | 67 | 740 | 10.9% | 16.9% | 89.9% | 72.5% | 8.3% | 3736 | 553 | 74.6% | 14.8% |
| [522, 529) | 777 | 47 | 730 | 7.6% | 16.6% | 97.6% | 89.1% | 6.0% | 4513 | 600 | 90.2% | 13.3% |
| [529, 539) | 493 | 15 | 478 | 2.4% | 10.9% | 100.0% | 100.0% | 3.0% | 5006 | 615 | 100.0% | 12.3% |



## 3.oot

表6 等宽分箱数据分布表

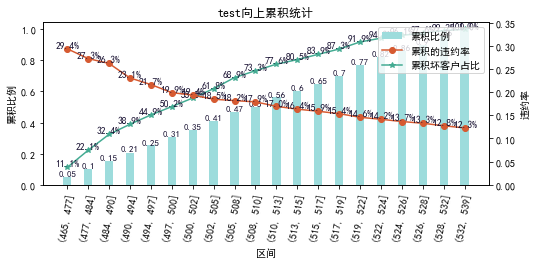
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [466, 473) | 25 | 11 | 14 | 3.2% | 0.8% | 3.2% | 0.8% | 44.0% | 25 | 11 | 1.2% | 44.0% |
| [473, 480) | 90 | 25 | 65 | 7.3% | 3.8% | 10.5% | 4.6% | 27.8% | 115 | 36 | 5.6% | 31.3% |
| [480, 487) | 97 | 33 | 64 | 9.6% | 3.7% | 20.2% | 8.3% | 34.0% | 212 | 69 | 10.3% | 32.5% |
| [487, 494) | 134 | 32 | 102 | 9.4% | 5.9% | 29.5% | 14.2% | 23.9% | 346 | 101 | 16.8% | 29.2% |
| [494, 501) | 226 | 57 | 169 | 16.7% | 9.8% | 46.2% | 24.0% | 25.2% | 572 | 158 | 27.7% | 27.6% |
| [501, 508) | 269 | 38 | 231 | 11.1% | 13.4% | 57.3% | 37.5% | 14.1% | 841 | 196 | 40.7% | 23.3% |
| [508, 515) | 298 | 40 | 258 | 11.7% | 15.0% | 69.0% | 52.4% | 13.4% | 1139 | 236 | 55.2% | 20.7% |
| [515, 522) | 339 | 46 | 293 | 13.5% | 17.0% | 82.5% | 69.5% | 13.6% | 1478 | 282 | 71.6% | 19.1% |
| [522, 529) | 343 | 40 | 303 | 11.7% | 17.6% | 94.2% | 87.0% | 11.7% | 1821 | 322 | 88.2% | 17.7% |
| [529, 539) | 243 | 20 | 223 | 5.8% | 13.0% | 100.0% | 100.0% | 8.2% | 2064 | 342 | 100.0% | 16.6% |

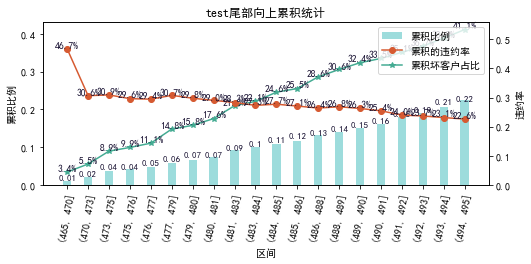


# 八、累积分数分布

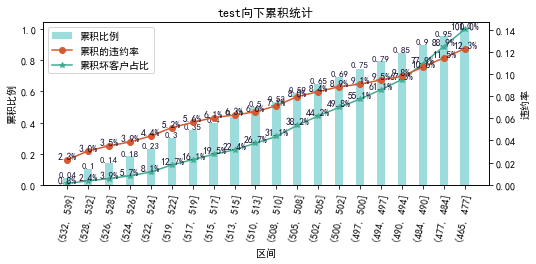
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

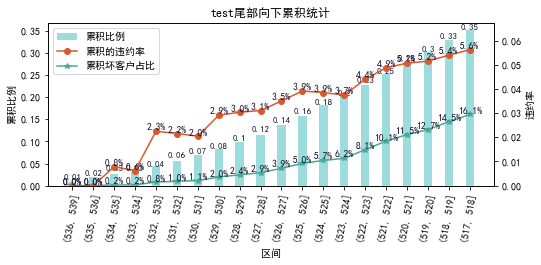
## 1. 向上累积分布





## 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 |
| [465.928, 473.2) | 2.5% | 2.2% | 1.4% | 0.0 | 0.006 |
| [473.2, 480.4) | 5.1% | 4.3% | 4.7% | 0.001 | 0.0 |
| [480.4, 487.6) | 5.8% | 5.9% | 4.9% | 0.0 | 0.001 |
| [487.6, 494.8) | 8.5% | 8.3% | 6.9% | 0.0 | 0.003 |
| [494.8, 502.0) | 11.9% | 12.5% | 11.3% | 0.0 | 0.0 |
| [502.0, 509.2) | 15.6% | 15.3% | 16.0% | 0.0 | 0.0 |
| [509.2, 516.4) | 14.5% | 13.7% | 13.7% | 0.0 | 0.0 |
| [516.4, 523.6) | 16.5% | 17.5% | 18.3% | 0.001 | 0.002 |
| [523.6, 530.8) | 13.2% | 13.4% | 14.4% | 0.0 | 0.001 |
| [530.8, 538.0) | 5.8% | 6.3% | 7.6% | 0.0 | 0.005 |
| sum | 99.4% | 99.4% | 99.2% | 0.004 | 0.019 |

# 十、策略建议