AKULAKU：id\_pdlold\_80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 | 2020-10-12 |

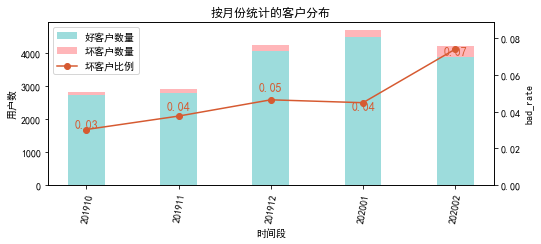
# 一、建模样本

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 |
|  | 18857 | 0.0484 | 201910 | 202002 | -- | -- |

表2 按月份统计分布表

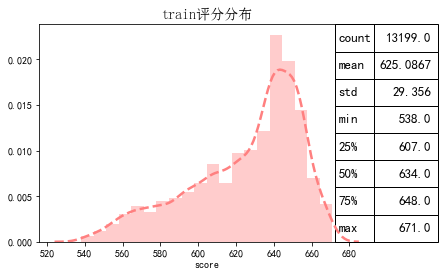
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 201910 | 2724 | 85 | 2809 | 3.0% | 1883 | 56 | 1939 | 2.9% | 841 | 29 | 870 | 3.3% |
| 201911 | 2792 | 109 | 2901 | 3.8% | 1935 | 82 | 2017 | 4.1% | 857 | 27 | 884 | 3.1% |
| 201912 | 4046 | 197 | 4243 | 4.6% | 2822 | 135 | 2957 | 4.6% | 1224 | 62 | 1286 | 4.8% |
| 202001 | 4492 | 211 | 4703 | 4.5% | 3176 | 144 | 3320 | 4.3% | 1316 | 67 | 1383 | 4.8% |
| 202002 | 3890 | 311 | 4201 | 7.4% | 2744 | 222 | 2966 | 7.5% | 1146 | 89 | 1235 | 7.2% |

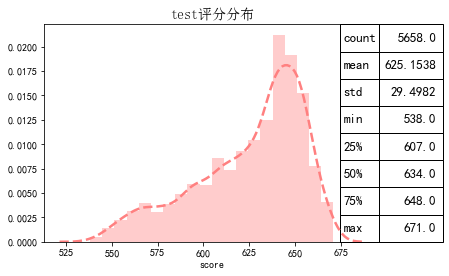


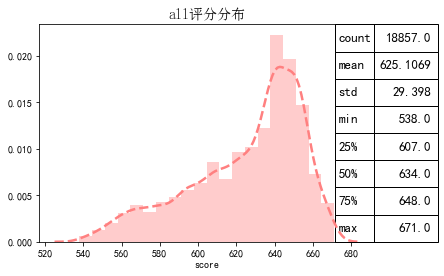
# 二、效果概况

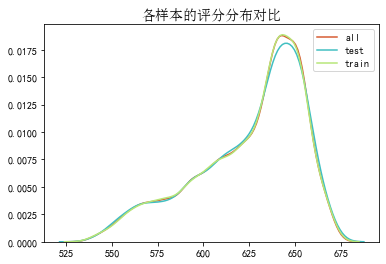
表3 模型效果概况表

|  |  |  |  |
| --- | --- | --- | --- |
|  | train | test | ALL |
| 样本量 | 13199 | 5658 | 18857 |
| 坏客户数量 | 639 | 274 | 913 |
| 坏客户比例 | 4.8% | 4.8% | 4.8% |
| KS | 0.264 | 0.239 | 0.254 |
| AUC | 0.67 | 0.649 | 0.664 |



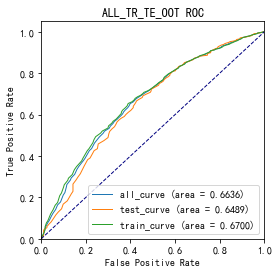
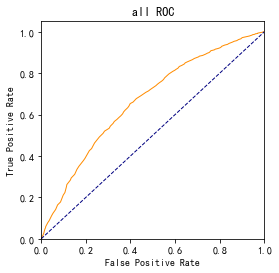
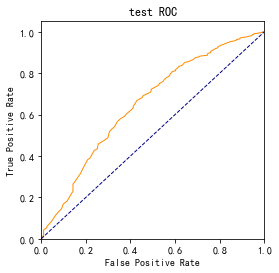
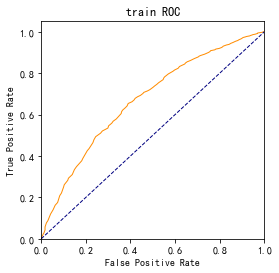






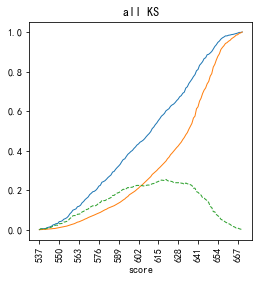
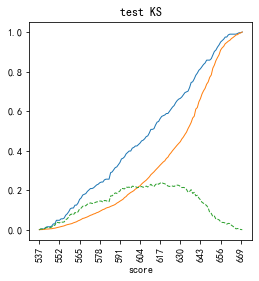
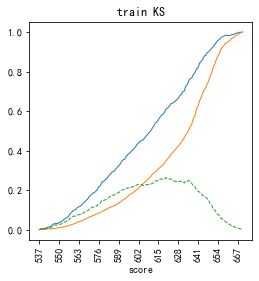
# 三、ROC曲线

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

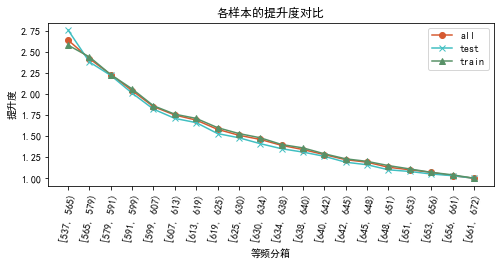
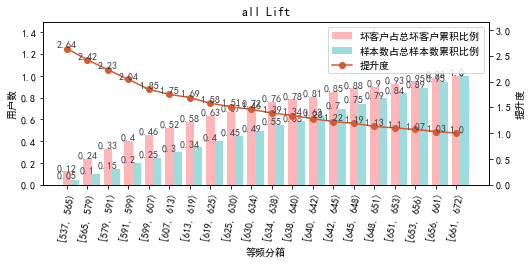
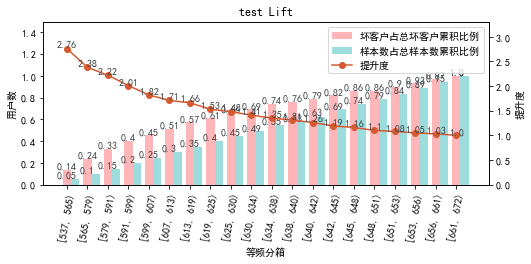
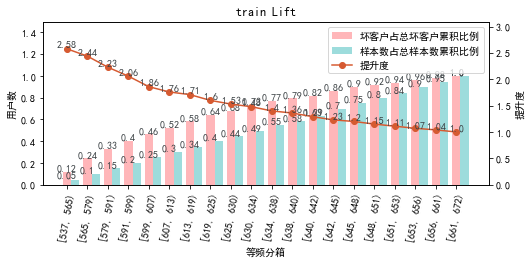
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

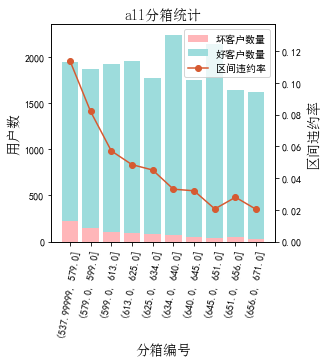
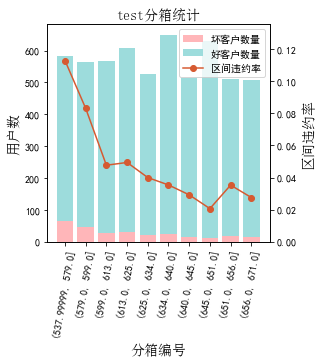
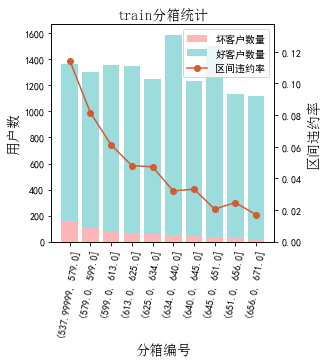
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



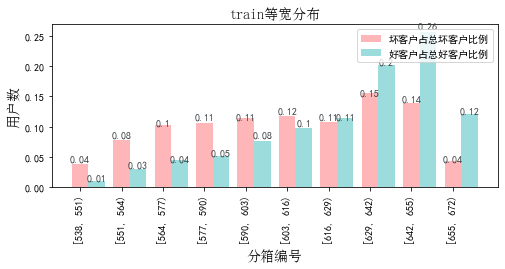
# 七、分数分布

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

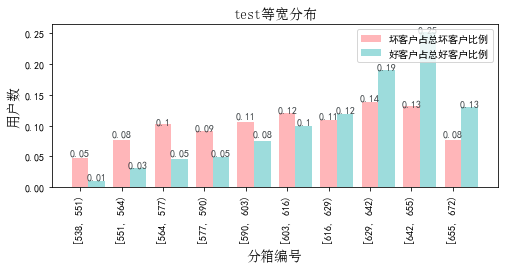
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [538, 551) | 156 | 24 | 132 | 3.8% | 1.1% | 3.8% | 1.1% | 15.4% | 156 | 24 | 1.2% | 15.4% |
| [551, 564) | 417 | 50 | 367 | 7.8% | 2.9% | 11.6% | 4.0% | 12.0% | 573 | 74 | 4.3% | 12.9% |
| [564, 577) | 618 | 65 | 553 | 10.2% | 4.4% | 21.8% | 8.4% | 10.5% | 1191 | 139 | 9.0% | 11.7% |
| [577, 590) | 709 | 68 | 641 | 10.6% | 5.1% | 32.4% | 13.5% | 9.6% | 1900 | 207 | 14.4% | 10.9% |
| [590, 603) | 1019 | 73 | 946 | 11.4% | 7.5% | 43.8% | 21.0% | 7.2% | 2919 | 280 | 22.1% | 9.6% |
| [603, 616) | 1299 | 75 | 1224 | 11.7% | 9.7% | 55.6% | 30.8% | 5.8% | 4218 | 355 | 32.0% | 8.4% |
| [616, 629) | 1497 | 69 | 1428 | 10.8% | 11.4% | 66.4% | 42.1% | 4.6% | 5715 | 424 | 43.3% | 7.4% |
| [629, 642) | 2637 | 99 | 2538 | 15.5% | 20.2% | 81.8% | 62.3% | 3.8% | 8352 | 523 | 63.3% | 6.3% |
| [642, 655) | 3310 | 89 | 3221 | 13.9% | 25.6% | 95.8% | 88.0% | 2.7% | 11662 | 612 | 88.4% | 5.2% |
| [655, 672) | 1537 | 27 | 1510 | 4.2% | 12.0% | 100.0% | 100.0% | 1.8% | 13199 | 639 | 100.0% | 4.8% |



## 2.test

表5 等宽分箱数据分布表

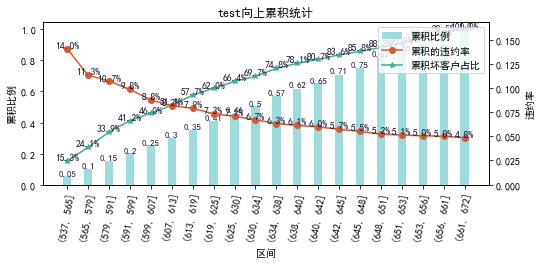
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [538, 551) | 64 | 13 | 51 | 4.7% | 0.9% | 4.7% | 0.9% | 20.3% | 64 | 13 | 1.1% | 20.3% |
| [551, 564) | 185 | 21 | 164 | 7.7% | 3.0% | 12.4% | 4.0% | 11.4% | 249 | 34 | 4.4% | 13.7% |
| [564, 577) | 275 | 28 | 247 | 10.2% | 4.6% | 22.6% | 8.6% | 10.2% | 524 | 62 | 9.3% | 11.8% |
| [577, 590) | 285 | 25 | 260 | 9.1% | 4.8% | 31.8% | 13.4% | 8.8% | 809 | 87 | 14.3% | 10.8% |
| [590, 603) | 436 | 29 | 407 | 10.6% | 7.6% | 42.3% | 21.0% | 6.7% | 1245 | 116 | 22.0% | 9.3% |
| [603, 616) | 569 | 33 | 536 | 12.0% | 10.0% | 54.4% | 30.9% | 5.8% | 1814 | 149 | 32.1% | 8.2% |
| [616, 629) | 669 | 30 | 639 | 10.9% | 11.9% | 65.3% | 42.8% | 4.5% | 2483 | 179 | 43.9% | 7.2% |
| [629, 642) | 1061 | 38 | 1023 | 13.9% | 19.0% | 79.2% | 61.8% | 3.6% | 3544 | 217 | 62.6% | 6.1% |
| [642, 655) | 1396 | 36 | 1360 | 13.1% | 25.3% | 92.3% | 87.1% | 2.6% | 4940 | 253 | 87.3% | 5.1% |
| [655, 672) | 718 | 21 | 697 | 7.7% | 12.9% | 100.0% | 100.0% | 2.9% | 5658 | 274 | 100.0% | 4.8% |

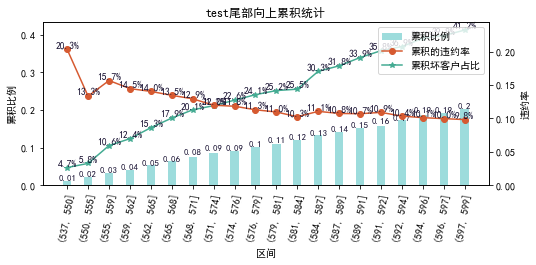


# 八、累积分数分布

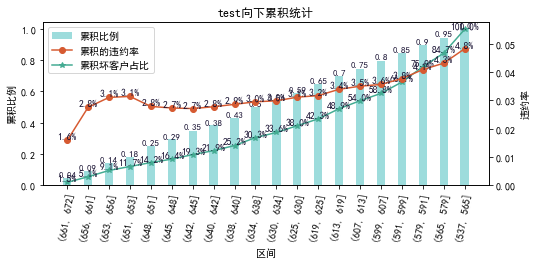
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

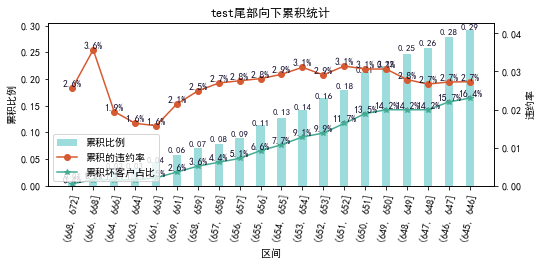
## 1. 向上累积分布





## 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 |
| [537.867, 551.3) | 1.3% | 1.3% | 0.0 |
| [551.3, 564.6) | 3.3% | 3.6% | 0.0 |
| [564.6, 577.9) | 4.8% | 4.6% | 0.0 |
| [577.9, 591.2) | 6.1% | 5.8% | 0.0 |
| [591.2, 604.5) | 7.9% | 7.8% | 0.0 |
| [604.5, 617.8) | 10.0% | 10.6% | 0.0 |
| [617.8, 631.1) | 13.2% | 13.1% | 0.0 |
| [631.1, 644.4) | 23.2% | 22.4% | 0.0 |
| [644.4, 657.7) | 22.8% | 22.9% | 0.0 |
| [657.7, 671.0) | 7.0% | 7.4% | 0.0 |
| sum | 99.6% | 99.6% | 0.001 |

# 十、策略建议