AKULAKU：idpdlold74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 | 2020-09-25 |

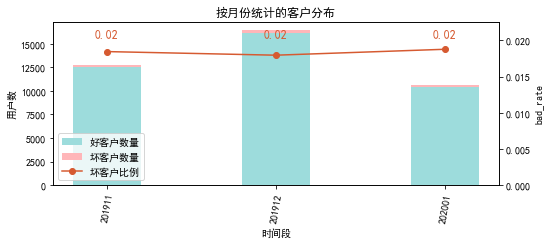
# 一、建模样本

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 |
|  | 39792 | 0.0183 | 201911 | 202001 | -- | -- |

表2 按月份统计分布表

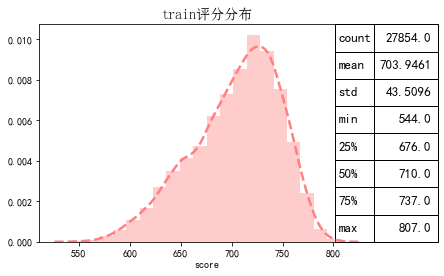
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 201911 | 12490 | 235 | 12725 | 1.8% | 8767 | 166 | 8933 | 1.9% | 3723 | 69 | 3792 | 1.8% |
| 201912 | 16178 | 296 | 16474 | 1.8% | 11251 | 205 | 11456 | 1.8% | 4927 | 91 | 5018 | 1.8% |
| 202001 | 10394 | 199 | 10593 | 1.9% | 7325 | 140 | 7465 | 1.9% | 3069 | 59 | 3128 | 1.9% |

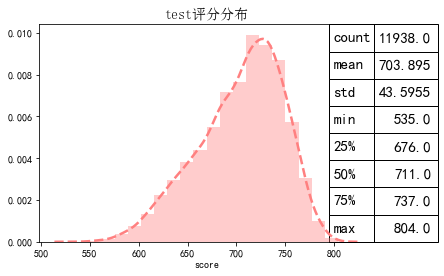


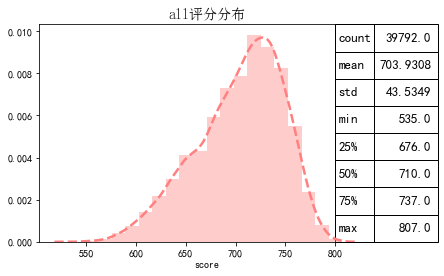
# 二、效果概况

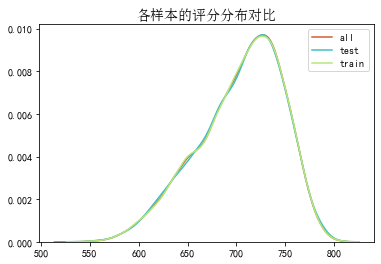
表3 模型效果概况表

|  |  |  |  |
| --- | --- | --- | --- |
|  | train | test | ALL |
| 样本量 | 27854 | 11938 | 39792 |
| 坏客户数量 | 511 | 219 | 730 |
| 坏客户比例 | 1.8% | 1.8% | 1.8% |
| KS | 0.365 | 0.365 | 0.362 |
| AUC | 0.75 | 0.73 | 0.744 |



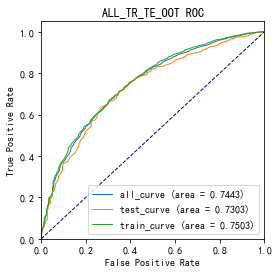
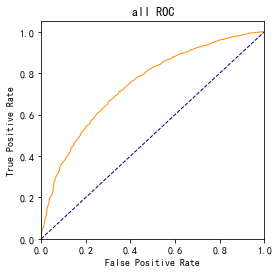
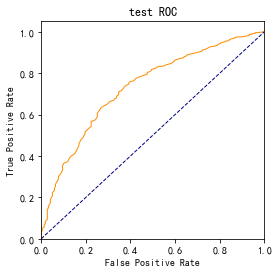
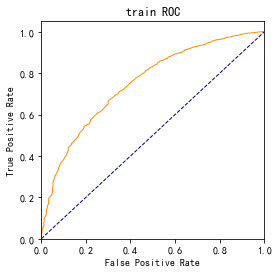






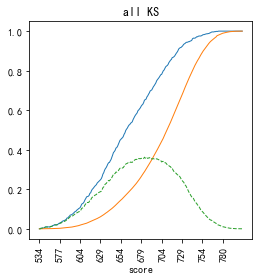
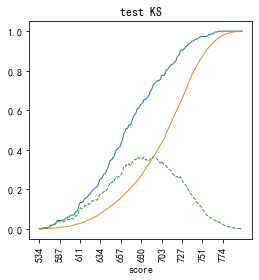
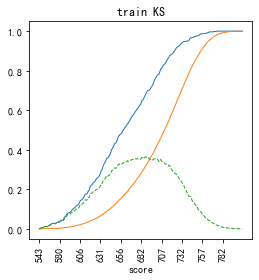
# 三、ROC曲线

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

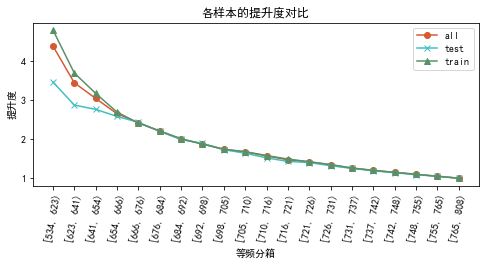
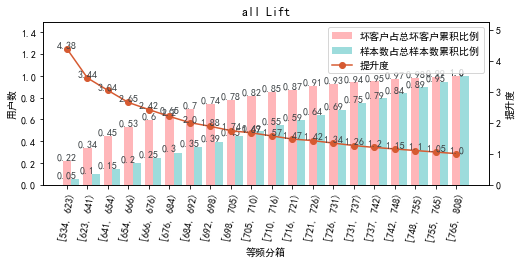
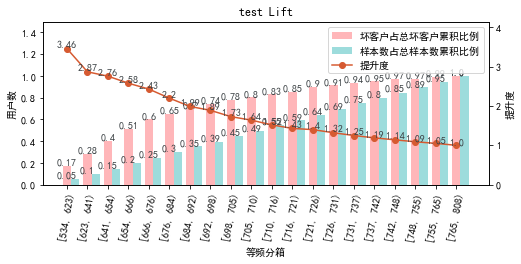
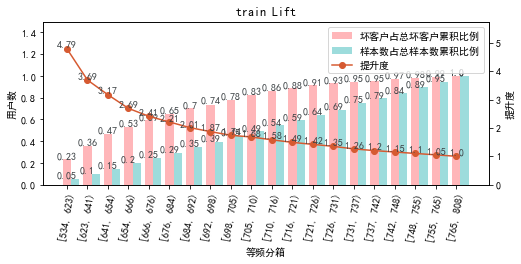
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

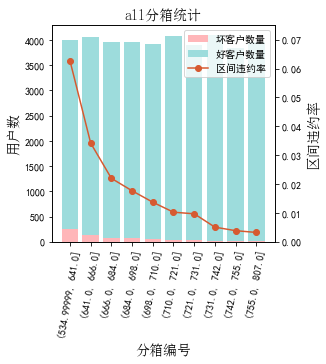
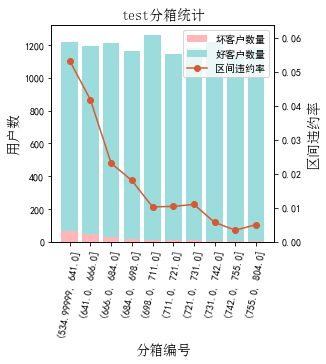
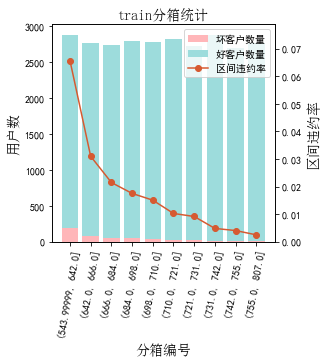
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



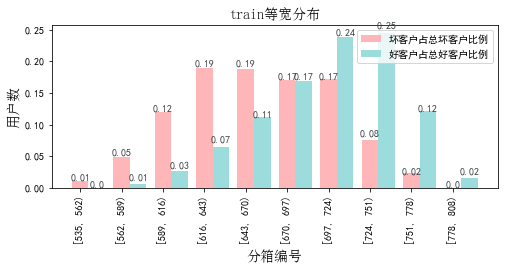
# 七、分数分布

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

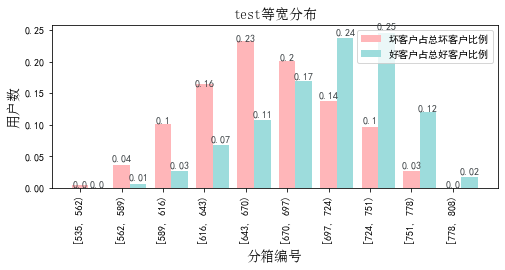
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [535, 562) | 18 | 6 | 12 | 1.2% | 0.0% | 1.2% | 0.0% | 33.3% | 18 | 6 | 0.1% | 33.3% |
| [562, 589) | 181 | 25 | 156 | 4.9% | 0.6% | 6.1% | 0.6% | 13.8% | 199 | 31 | 0.7% | 15.6% |
| [589, 616) | 802 | 61 | 741 | 11.9% | 2.7% | 18.0% | 3.3% | 7.6% | 1001 | 92 | 3.6% | 9.2% |
| [616, 643) | 1876 | 97 | 1779 | 19.0% | 6.5% | 37.0% | 9.8% | 5.2% | 2877 | 189 | 10.3% | 6.6% |
| [643, 670) | 3155 | 96 | 3059 | 18.8% | 11.2% | 55.8% | 21.0% | 3.0% | 6032 | 285 | 21.7% | 4.7% |
| [670, 697) | 4708 | 87 | 4621 | 17.0% | 16.9% | 72.8% | 37.9% | 1.8% | 10740 | 372 | 38.6% | 3.5% |
| [697, 724) | 6628 | 88 | 6540 | 17.2% | 23.9% | 90.0% | 61.8% | 1.3% | 17368 | 460 | 62.4% | 2.6% |
| [724, 751) | 6747 | 39 | 6708 | 7.6% | 24.5% | 97.7% | 86.4% | 0.6% | 24115 | 499 | 86.6% | 2.1% |
| [751, 778) | 3320 | 12 | 3308 | 2.3% | 12.1% | 100.0% | 98.5% | 0.4% | 27435 | 511 | 98.5% | 1.9% |
| [778, 808) | 419 | 0 | 419 | 0.0% | 1.5% | 100.0% | 100.0% | 0.0% | 27854 | 511 | 100.0% | 1.8% |



## 2.test

表5 等宽分箱数据分布表

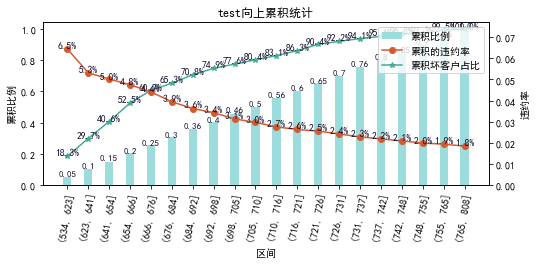
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [535, 562) | 9 | 1 | 8 | 0.5% | 0.1% | 0.5% | 0.1% | 11.1% | 9 | 1 | 0.1% | 11.1% |
| [562, 589) | 79 | 8 | 71 | 3.7% | 0.6% | 4.1% | 0.7% | 10.1% | 88 | 9 | 0.7% | 10.2% |
| [589, 616) | 333 | 22 | 311 | 10.0% | 2.7% | 14.2% | 3.3% | 6.6% | 421 | 31 | 3.5% | 7.4% |
| [616, 643) | 839 | 36 | 803 | 16.4% | 6.9% | 30.6% | 10.2% | 4.3% | 1260 | 67 | 10.6% | 5.3% |
| [643, 670) | 1320 | 51 | 1269 | 23.3% | 10.8% | 53.9% | 21.0% | 3.9% | 2580 | 118 | 21.6% | 4.6% |
| [670, 697) | 2034 | 44 | 1990 | 20.1% | 17.0% | 74.0% | 38.0% | 2.2% | 4614 | 162 | 38.6% | 3.5% |
| [697, 724) | 2820 | 30 | 2790 | 13.7% | 23.8% | 87.7% | 61.8% | 1.1% | 7434 | 192 | 62.3% | 2.6% |
| [724, 751) | 2903 | 21 | 2882 | 9.6% | 24.6% | 97.3% | 86.4% | 0.7% | 10337 | 213 | 86.6% | 2.1% |
| [751, 778) | 1407 | 6 | 1401 | 2.7% | 12.0% | 100.0% | 98.3% | 0.4% | 11744 | 219 | 98.4% | 1.9% |
| [778, 808) | 194 | 0 | 194 | 0.0% | 1.7% | 100.0% | 100.0% | 0.0% | 11938 | 219 | 100.0% | 1.8% |

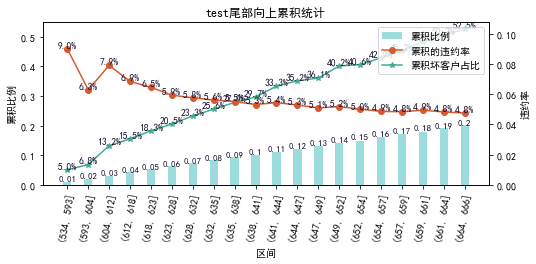


# 八、累积分数分布

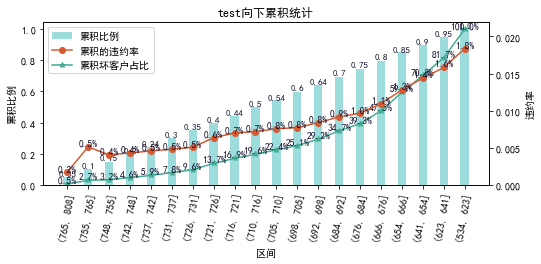
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

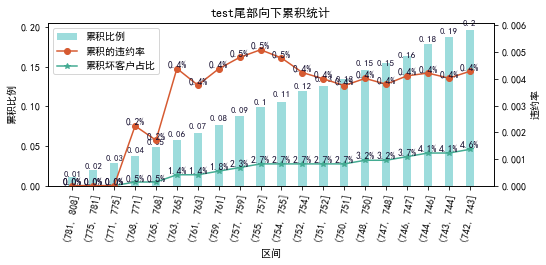
## 1. 向上累积分布





## 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 |
| [534.728, 562.2) | 0.1% | 0.1% | 0.0 |
| [562.2, 589.4) | 0.7% | 0.7% | 0.0 |
| [589.4, 616.6) | 3.0% | 3.0% | 0.0 |
| [616.6, 643.8) | 6.9% | 7.1% | 0.0 |
| [643.8, 671.0) | 11.5% | 11.3% | 0.0 |
| [671.0, 698.2) | 18.0% | 18.0% | 0.0 |
| [698.2, 725.4) | 24.2% | 23.9% | 0.0 |
| [725.4, 752.6) | 23.7% | 24.1% | 0.0 |
| [752.6, 779.8) | 10.8% | 10.5% | 0.0 |
| [779.8, 807.0) | 1.2% | 1.3% | 0.0 |
| sum | 100.0% | 100.0% | 0.0 |

# 十、策略建议