AKULAKU：idantifraud10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 | 2020-08-28 |

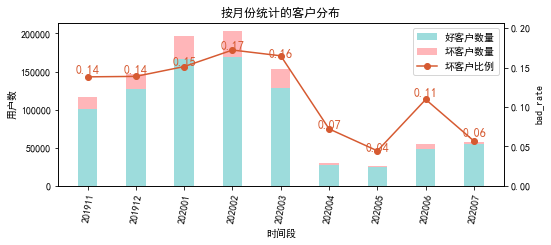
# 一、建模样本

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 |
|  | 987337 | 0.1412 | 201911 | 202001 | 202002 | 202007 |

表2 按月份统计分布表

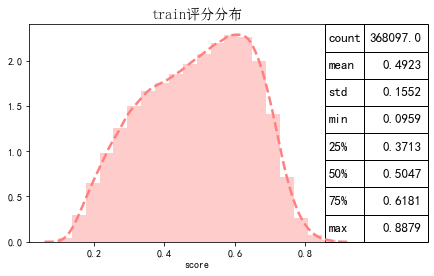
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 201911 | 100423.0 | 16109.0 | 116532 | 13.8% | 80338.0 | 12856.0 | 93194.0 | 13.8% | 20085.0 | 3253.0 | 23338.0 | 13.9% |
| 201912 | 126782.0 | 20437.0 | 147219 | 13.9% | 101385.0 | 16434.0 | 117819.0 | 13.9% | 25397.0 | 4003.0 | 29400.0 | 13.6% |
| 202001 | 166729.0 | 29711.0 | 196440 | 15.1% | 133370.0 | 23714.0 | 157084.0 | 15.1% | 33359.0 | 5997.0 | 39356.0 | 15.2% |
| 202002 | 168684.0 | 35094.0 | 203778 | 17.2% | 0.0 | 0.0 | 0.0 | 0.0% | 0.0 | 0.0 | 0.0 | 0.0% |
| 202003 | 128360.0 | 25346.0 | 153706 | 16.5% | 0.0 | 0.0 | 0.0 | 0.0% | 0.0 | 0.0 | 0.0 | 0.0% |
| 202004 | 28078.0 | 2187.0 | 30265 | 7.2% | 0.0 | 0.0 | 0.0 | 0.0% | 0.0 | 0.0 | 0.0 | 0.0% |
| 202005 | 25208.0 | 1173.0 | 26381 | 4.4% | 0.0 | 0.0 | 0.0 | 0.0% | 0.0 | 0.0 | 0.0 | 0.0% |
| 202006 | 48912.0 | 6020.0 | 54932 | 11.0% | 0.0 | 0.0 | 0.0 | 0.0% | 0.0 | 0.0 | 0.0 | 0.0% |
| 202007 | 54784.0 | 3300.0 | 58084 | 5.7% | 0.0 | 0.0 | 0.0 | 0.0% | 0.0 | 0.0 | 0.0 | 0.0% |

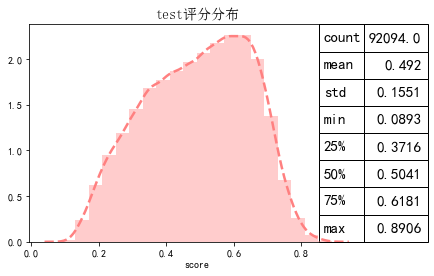


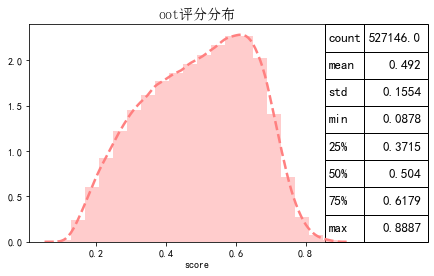
# 二、效果概况

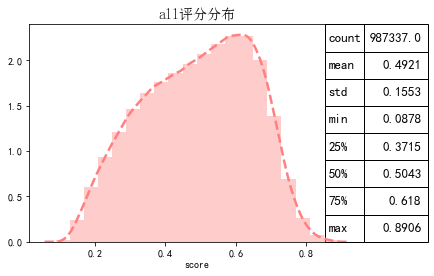
表3 模型效果概况表

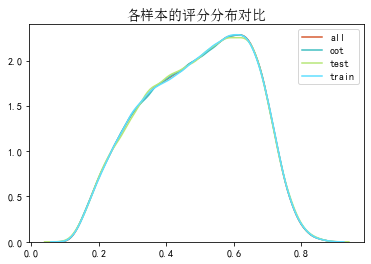
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | train | test | oot | ALL |
| 样本量 | 368097 | 92094 | 527146 | 987337 |
| 坏客户数量 | 53004 | 13253 | 73120 | 139377 |
| 坏客户比例 | 14.4% | 14.4% | 13.9% | 14.1% |
| KS | 0.00368 | 0.00436 | 0.00526 | 0.00424 |
| AUC | 0.501 | 0.501 | 0.502 | 0.501 |





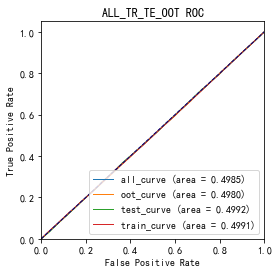
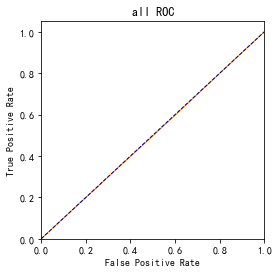
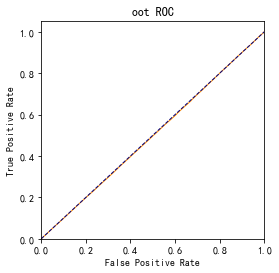
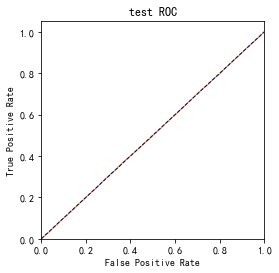
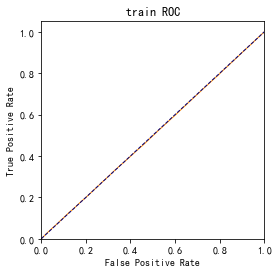






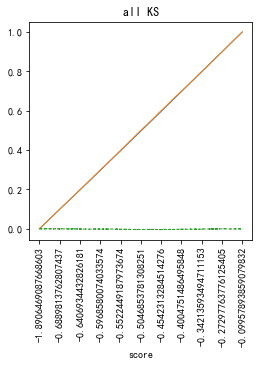
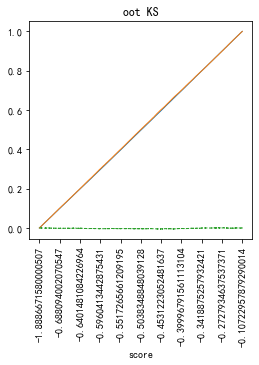
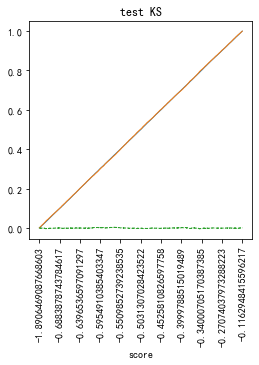
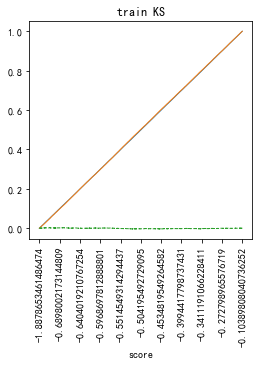
# 三、ROC曲线

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

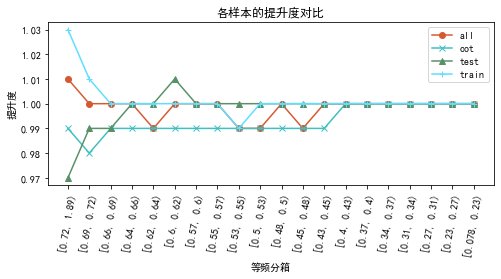
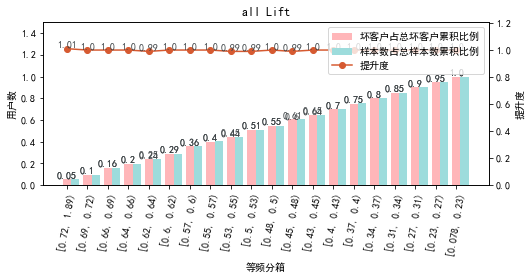
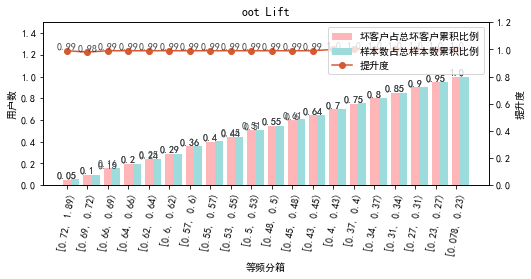
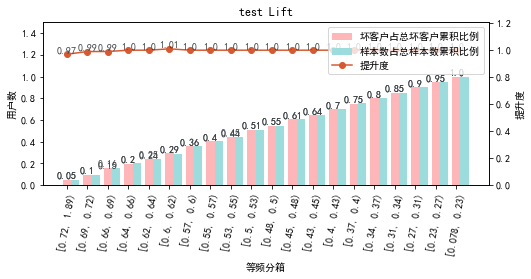
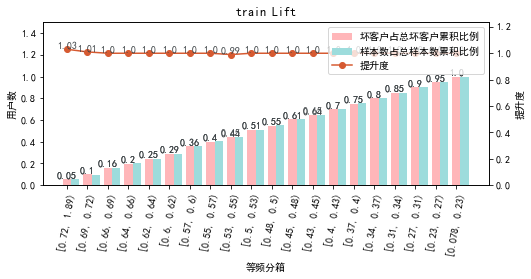
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

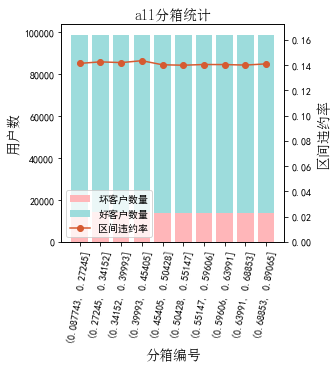
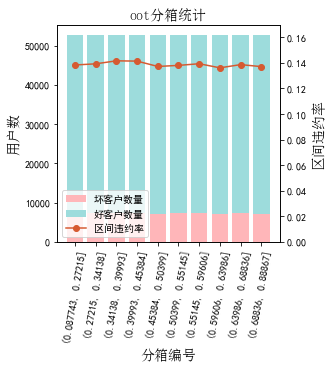
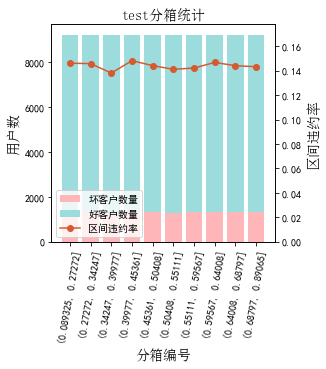
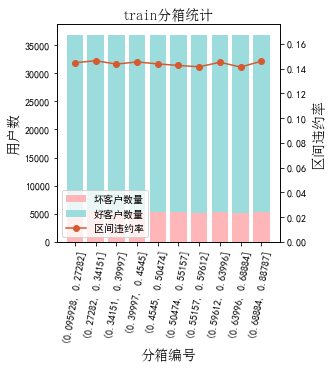
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



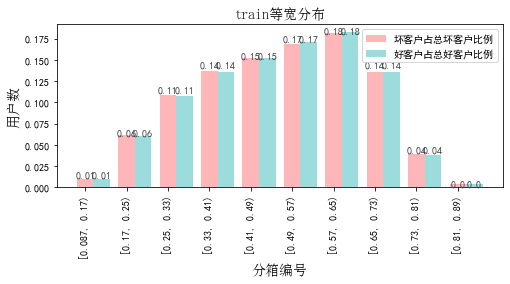
# 七、分数分布

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

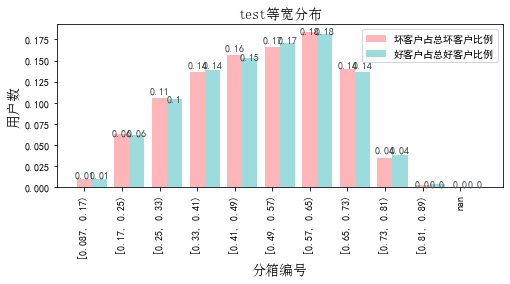
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计通过率 | 累计拒绝坏人占比 |
| [0.087, 0.17) | 3809 | 550 | 3259 | 1.0% | 1.0% | 1.0% | 1.0% | 14.4% | 3809 | 550 | 1.0% | 14.4% |
| [0.17, 0.25) | 22496 | 3254 | 19242 | 6.1% | 6.1% | 7.2% | 7.1% | 14.5% | 26305 | 3804 | 7.1% | 14.5% |
| [0.25, 0.33) | 39766 | 5781 | 33985 | 10.9% | 10.8% | 18.1% | 17.9% | 14.5% | 66071 | 9585 | 17.9% | 14.5% |
| [0.33, 0.41) | 50143 | 7244 | 42899 | 13.7% | 13.6% | 31.8% | 31.5% | 14.4% | 116214 | 16829 | 31.6% | 14.5% |
| [0.41, 0.49) | 56153 | 8091 | 48062 | 15.3% | 15.3% | 47.0% | 46.8% | 14.4% | 172367 | 24920 | 46.8% | 14.5% |
| [0.49, 0.57) | 62988 | 8972 | 54016 | 16.9% | 17.1% | 63.9% | 63.9% | 14.2% | 235355 | 33892 | 63.9% | 14.4% |
| [0.57, 0.65) | 67205 | 9635 | 57570 | 18.2% | 18.3% | 82.1% | 82.2% | 14.3% | 302560 | 43527 | 82.2% | 14.4% |
| [0.65, 0.73) | 50132 | 7181 | 42951 | 13.5% | 13.6% | 95.7% | 95.8% | 14.3% | 352692 | 50708 | 95.8% | 14.4% |
| [0.73, 0.81) | 14108 | 2109 | 11999 | 4.0% | 3.8% | 99.6% | 99.6% | 14.9% | 366800 | 52817 | 99.6% | 14.4% |
| [0.81, 0.89) | 1297 | 187 | 1110 | 0.4% | 0.4% | 100.0% | 100.0% | 14.4% | 368097 | 53004 | 100.0% | 14.4% |



## 2.test

表5 等宽分箱数据分布表

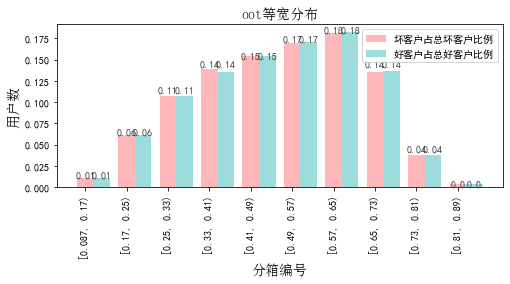
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计通过率 | 累计拒绝坏人占比 |
| [0.087, 0.17) | 942 | 137 | 805 | 1.0% | 1.0% | 1.0% | 1.0% | 14.5% | 942 | 137 | 1.0% | 14.5% |
| [0.17, 0.25) | 5740 | 842 | 4898 | 6.4% | 6.2% | 7.4% | 7.2% | 14.7% | 6682 | 979 | 7.3% | 14.7% |
| [0.25, 0.33) | 9678 | 1400 | 8278 | 10.6% | 10.5% | 18.0% | 17.7% | 14.5% | 16360 | 2379 | 17.8% | 14.5% |
| [0.33, 0.41) | 12764 | 1808 | 10956 | 13.6% | 13.9% | 31.6% | 31.6% | 14.2% | 29124 | 4187 | 31.6% | 14.4% |
| [0.41, 0.49) | 14176 | 2076 | 12100 | 15.7% | 15.3% | 47.3% | 47.0% | 14.6% | 43300 | 6263 | 47.0% | 14.5% |
| [0.49, 0.57) | 15697 | 2199 | 13498 | 16.6% | 17.1% | 63.8% | 64.1% | 14.0% | 58997 | 8462 | 64.1% | 14.3% |
| [0.57, 0.65) | 16723 | 2433 | 14290 | 18.4% | 18.1% | 82.2% | 82.2% | 14.5% | 75720 | 10895 | 82.2% | 14.4% |
| [0.65, 0.73) | 12584 | 1850 | 10734 | 14.0% | 13.6% | 96.2% | 95.8% | 14.7% | 88304 | 12745 | 95.9% | 14.4% |
| [0.73, 0.81) | 3467 | 468 | 2999 | 3.5% | 3.8% | 99.7% | 99.6% | 13.5% | 91771 | 13213 | 99.6% | 14.4% |
| [0.81, 0.89) | 322 | 39 | 283 | 0.3% | 0.4% | 100.0% | 100.0% | 12.1% | 92093 | 13252 | 100.0% | 14.4% |
| nan | 1 | 1 | 0 | 0.0% | 0.0% | 100.0% | 100.0% | 100.0% | 92094 | 13253 | 100.0% | 14.4% |



## 3.oot

表6 等宽分箱数据分布表

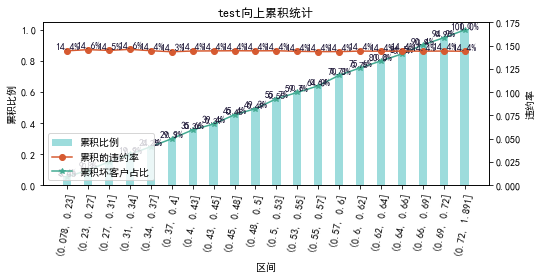
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计通过率 | 累计拒绝坏人占比 |
| [0.087, 0.17) | 5602 | 795 | 4807 | 1.1% | 1.1% | 1.1% | 1.1% | 14.2% | 5602 | 795 | 1.1% | 14.2% |
| [0.17, 0.25) | 32495 | 4498 | 27997 | 6.2% | 6.2% | 7.2% | 7.2% | 13.8% | 38097 | 5293 | 7.2% | 13.9% |
| [0.25, 0.33) | 56767 | 7859 | 48908 | 10.7% | 10.8% | 18.0% | 18.0% | 13.8% | 94864 | 13152 | 18.0% | 13.9% |
| [0.33, 0.41) | 71864 | 10183 | 61681 | 13.9% | 13.6% | 31.9% | 31.6% | 14.2% | 166728 | 23335 | 31.6% | 14.0% |
| [0.41, 0.49) | 80987 | 11260 | 69727 | 15.4% | 15.4% | 47.3% | 46.9% | 13.9% | 247715 | 34595 | 47.0% | 14.0% |
| [0.49, 0.57) | 89671 | 12368 | 77303 | 16.9% | 17.0% | 64.2% | 64.0% | 13.8% | 337386 | 46963 | 64.0% | 13.9% |
| [0.57, 0.65) | 95904 | 13207 | 82697 | 18.1% | 18.2% | 82.3% | 82.2% | 13.8% | 433290 | 60170 | 82.2% | 13.9% |
| [0.65, 0.73) | 71860 | 9913 | 61947 | 13.6% | 13.6% | 95.8% | 95.8% | 13.8% | 505150 | 70083 | 95.8% | 13.9% |
| [0.73, 0.81) | 20154 | 2777 | 17377 | 3.8% | 3.8% | 99.6% | 99.7% | 13.8% | 525304 | 72860 | 99.7% | 13.9% |
| [0.81, 0.89) | 1842 | 260 | 1582 | 0.4% | 0.3% | 100.0% | 100.0% | 14.1% | 527146 | 73120 | 100.0% | 13.9% |

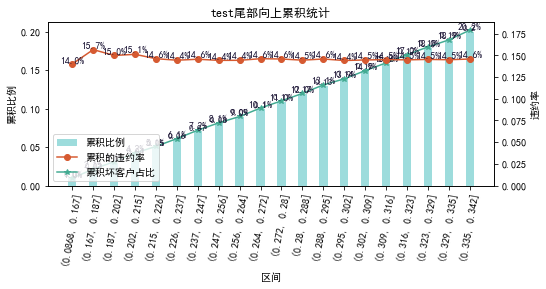


# 八、累积分数分布

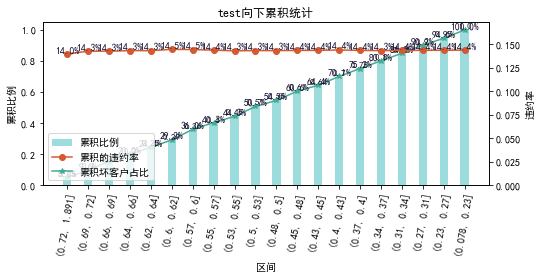
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

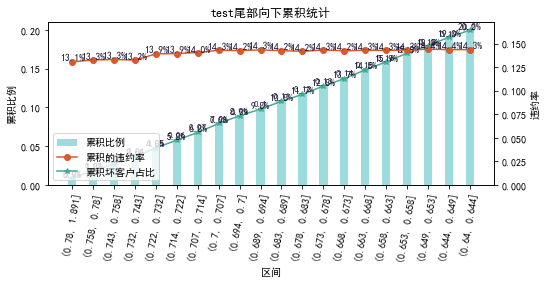
## 1. 向上累积分布





## 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 |
| [0.0869, 0.168) | 1.0% | 1.0% | 1.1% | 0.0 | 0.0 |
| [0.168, 0.248) | 6.1% | 6.2% | 6.2% | 0.0 | 0.0 |
| [0.248, 0.329) | 10.8% | 10.5% | 10.8% | 0.0 | 0.0 |
| [0.329, 0.409) | 13.6% | 13.9% | 13.6% | 0.0 | 0.0 |
| [0.409, 0.489) | 15.3% | 15.4% | 15.4% | 0.0 | 0.0 |
| [0.489, 0.569) | 17.1% | 17.0% | 17.0% | 0.0 | 0.0 |
| [0.569, 0.65) | 18.3% | 18.2% | 18.2% | 0.0 | 0.0 |
| [0.65, 0.73) | 13.6% | 13.7% | 13.6% | 0.0 | 0.0 |
| [0.73, 0.81) | 3.8% | 3.8% | 3.8% | 0.0 | 0.0 |
| [0.81, 0.891) | 0.4% | 0.3% | 0.3% | 0.0 | 0.0 |
| sum | 100.0% | 100.0% | 100.0% | 0.0 | 0.0 |

# 十、策略建议