AKULAKU：antifraud10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 |

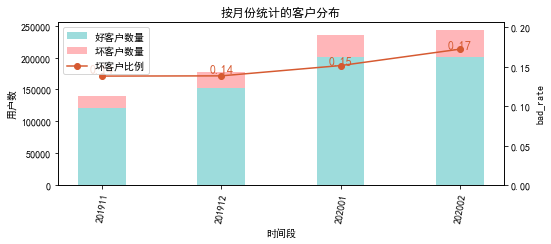
# 一、建模样本

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 |
|  | 797039 | 0.1526 | 201911 | 202001 | 202002 | 202002 |

表2 按月份统计分布表

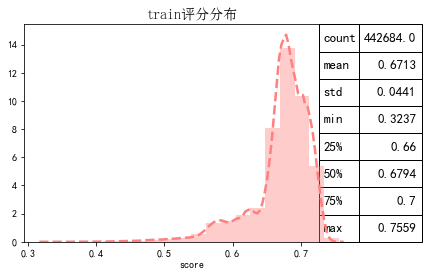
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 201911 | 121078 | 19412 | 140490 | 13.8% | 96906.0 | 15593.0 | 112499.0 | 13.9% | 24172.0 | 3819.0 | 27991.0 | 13.6% |
| 201912 | 152360 | 24487 | 176847 | 13.8% | 121840.0 | 19483.0 | 141323.0 | 13.8% | 30520.0 | 5004.0 | 35524.0 | 14.1% |
| 202001 | 200255 | 35763 | 236018 | 15.2% | 160121.0 | 28741.0 | 188862.0 | 15.2% | 40134.0 | 7022.0 | 47156.0 | 14.9% |
| 202002 | 201689 | 41995 | 243684 | 17.2% | 0.0 | 0.0 | 0.0 | 0.0% | 0.0 | 0.0 | 0.0 | 0.0% |

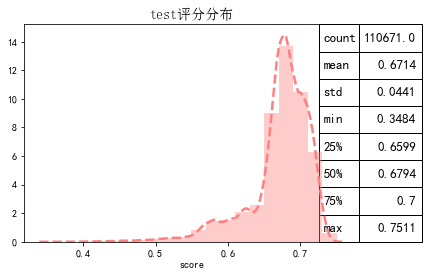


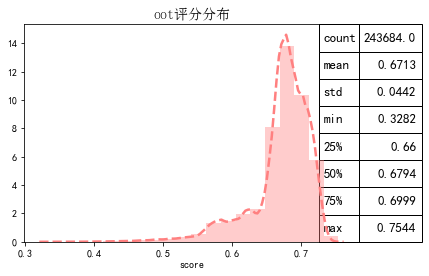
# 二、效果概况

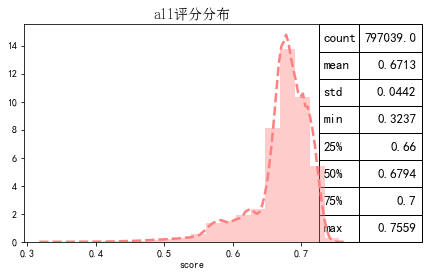
表3 模型效果概况表

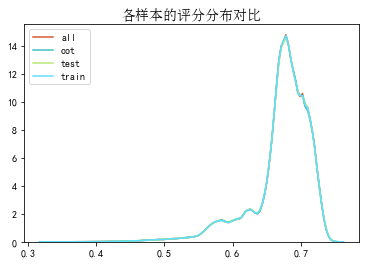
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | train | test | oot | ALL |
| 样本量 | 442684 | 110671 | 243684 | 797039 |
| 坏客户数量 | 63817 | 15845 | 41995 | 121657 |
| 坏客户比例 | 14.4% | 14.3% | 17.2% | 15.3% |
| KS | 0.0038 | 0.00647 | 0.00439 | 0.00255 |
| AUC | 0.499 | 0.5 | 0.502 | 0.5 |





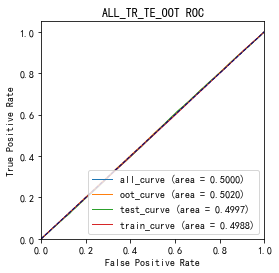
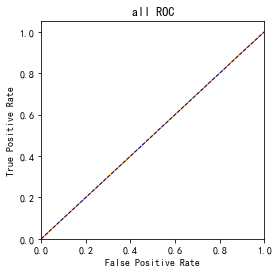
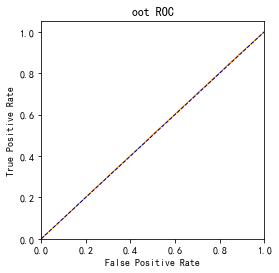
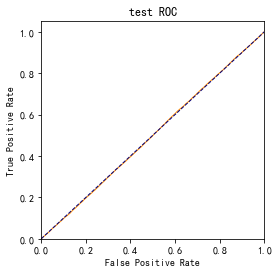
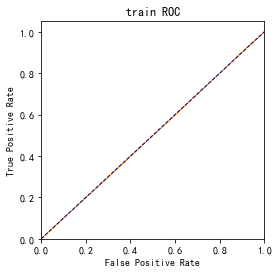






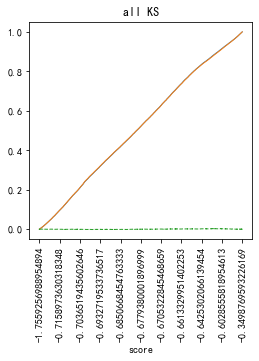
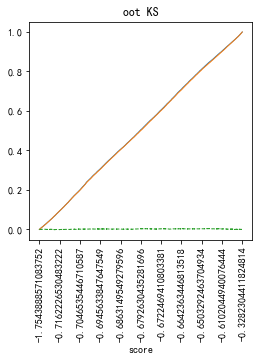
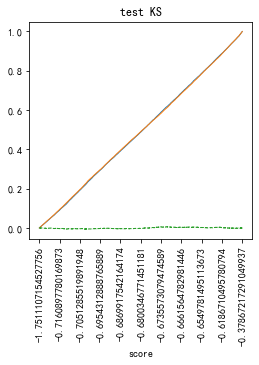
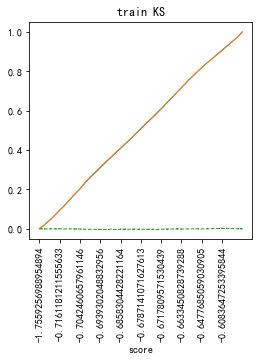
# 三、ROC曲线

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

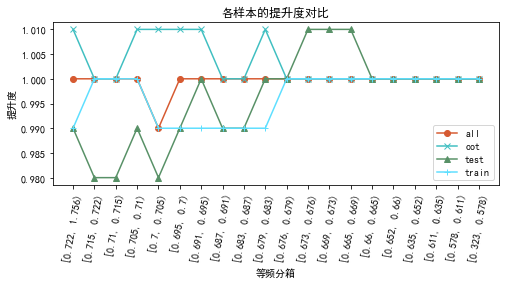
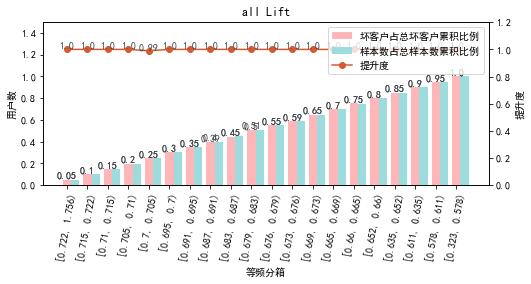
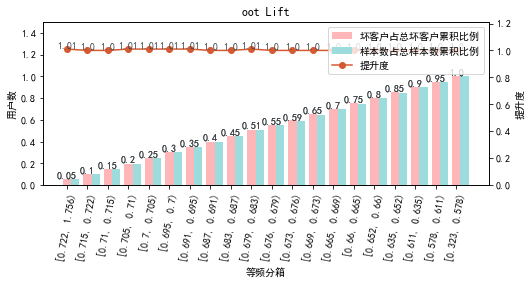
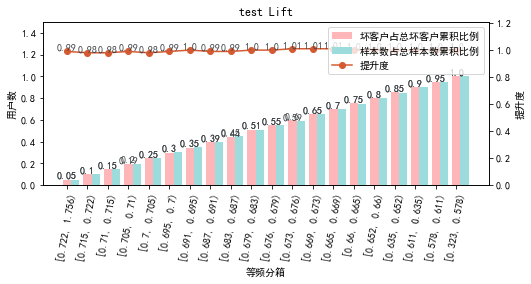
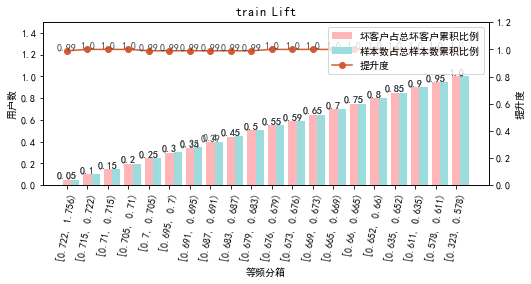
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

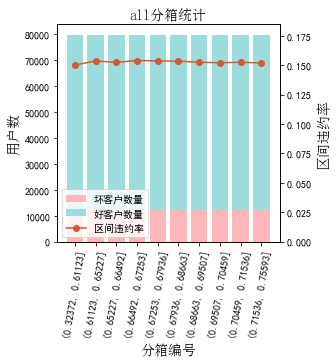
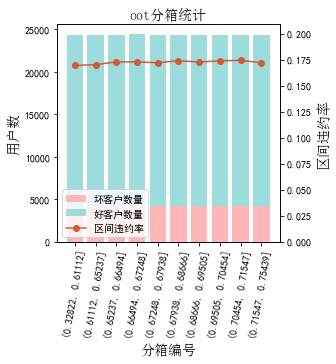
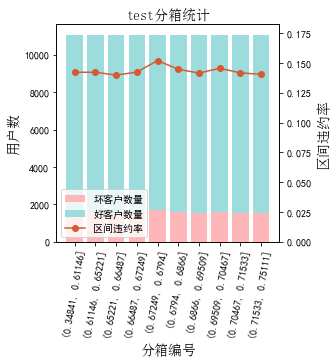
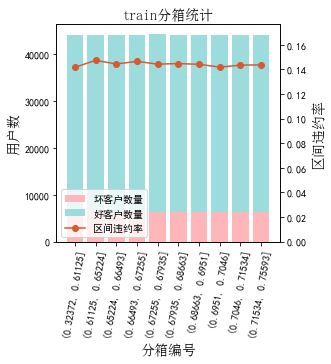
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



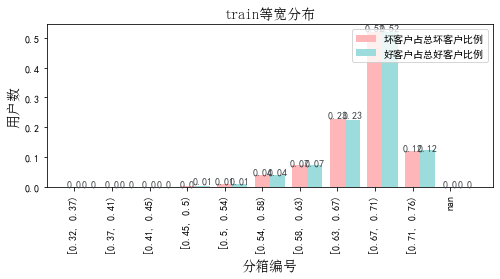
# 七、分数分布

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

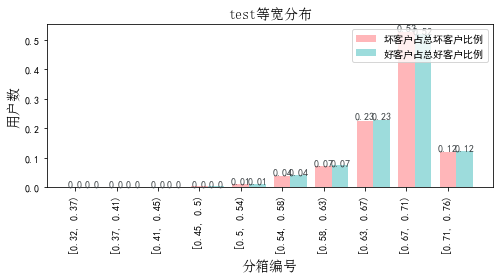
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计通过率 | 累计拒绝坏人占比 |
| [0.32, 0.37) | 17 | 3 | 14 | 0.0% | 0.0% | 0.0% | 0.0% | 17.6% | 17 | 3 | 0.0% | 17.6% |
| [0.37, 0.41) | 148 | 25 | 123 | 0.0% | 0.0% | 0.0% | 0.0% | 16.9% | 165 | 28 | 0.0% | 17.0% |
| [0.41, 0.45) | 698 | 101 | 597 | 0.2% | 0.2% | 0.2% | 0.2% | 14.5% | 863 | 129 | 0.2% | 14.9% |
| [0.45, 0.5) | 2214 | 318 | 1896 | 0.5% | 0.5% | 0.7% | 0.7% | 14.4% | 3077 | 447 | 0.7% | 14.5% |
| [0.5, 0.54) | 4619 | 650 | 3969 | 1.0% | 1.0% | 1.7% | 1.7% | 14.1% | 7696 | 1097 | 1.7% | 14.3% |
| [0.54, 0.58) | 17776 | 2501 | 15275 | 3.9% | 4.0% | 5.6% | 5.8% | 14.1% | 25472 | 3598 | 5.8% | 14.1% |
| [0.58, 0.63) | 32446 | 4684 | 27762 | 7.3% | 7.3% | 13.0% | 13.1% | 14.4% | 57918 | 8282 | 13.1% | 14.3% |
| [0.63, 0.67) | 100275 | 14635 | 85640 | 22.9% | 22.6% | 35.9% | 35.7% | 14.6% | 158193 | 22917 | 35.7% | 14.5% |
| [0.67, 0.71) | 230142 | 33118 | 197024 | 51.9% | 52.0% | 87.8% | 87.7% | 14.4% | 388335 | 56035 | 87.7% | 14.4% |
| [0.71, 0.76) | 54348 | 7781 | 46567 | 12.2% | 12.3% | 100.0% | 100.0% | 14.3% | 442683 | 63816 | 100.0% | 14.4% |
| nan | 1 | 1 | 0 | 0.0% | 0.0% | 100.0% | 100.0% | 100.0% | 442684 | 63817 | 100.0% | 14.4% |



## 2.test

表5 等宽分箱数据分布表

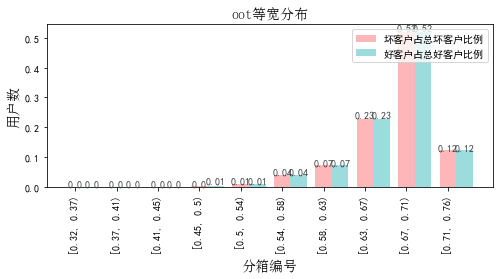
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计通过率 | 累计拒绝坏人占比 |
| [0.32, 0.37) | 6 | 1 | 5 | 0.0% | 0.0% | 0.0% | 0.0% | 16.7% | 6 | 1 | 0.0% | 16.7% |
| [0.37, 0.41) | 46 | 7 | 39 | 0.0% | 0.0% | 0.1% | 0.0% | 15.2% | 52 | 8 | 0.0% | 15.4% |
| [0.41, 0.45) | 191 | 21 | 170 | 0.1% | 0.2% | 0.2% | 0.2% | 11.0% | 243 | 29 | 0.2% | 11.9% |
| [0.45, 0.5) | 501 | 75 | 426 | 0.5% | 0.4% | 0.7% | 0.7% | 15.0% | 744 | 104 | 0.7% | 14.0% |
| [0.5, 0.54) | 1164 | 191 | 973 | 1.2% | 1.0% | 1.9% | 1.7% | 16.4% | 1908 | 295 | 1.7% | 15.5% |
| [0.54, 0.58) | 4416 | 623 | 3793 | 3.9% | 4.0% | 5.8% | 5.7% | 14.1% | 6324 | 918 | 5.7% | 14.5% |
| [0.58, 0.63) | 8159 | 1115 | 7044 | 7.0% | 7.4% | 12.8% | 13.1% | 13.7% | 14483 | 2033 | 13.1% | 14.0% |
| [0.63, 0.67) | 25045 | 3576 | 21469 | 22.6% | 22.6% | 35.4% | 35.8% | 14.3% | 39528 | 5609 | 35.7% | 14.2% |
| [0.67, 0.71) | 57629 | 8370 | 49259 | 52.8% | 51.9% | 88.2% | 87.7% | 14.5% | 97157 | 13979 | 87.8% | 14.4% |
| [0.71, 0.76) | 13514 | 1866 | 11648 | 11.8% | 12.3% | 100.0% | 100.0% | 13.8% | 110671 | 15845 | 100.0% | 14.3% |



## 3.oot

表6 等宽分箱数据分布表

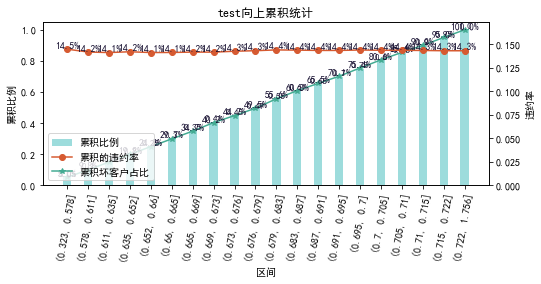
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计通过率 | 累计拒绝坏人占比 |
| [0.32, 0.37) | 9 | 3 | 6 | 0.0% | 0.0% | 0.0% | 0.0% | 33.3% | 9 | 3 | 0.0% | 33.3% |
| [0.37, 0.41) | 92 | 19 | 73 | 0.0% | 0.0% | 0.1% | 0.0% | 20.7% | 101 | 22 | 0.0% | 21.8% |
| [0.41, 0.45) | 399 | 66 | 333 | 0.2% | 0.2% | 0.2% | 0.2% | 16.5% | 500 | 88 | 0.2% | 17.6% |
| [0.45, 0.5) | 1220 | 203 | 1017 | 0.5% | 0.5% | 0.7% | 0.7% | 16.6% | 1720 | 291 | 0.7% | 16.9% |
| [0.5, 0.54) | 2609 | 445 | 2164 | 1.1% | 1.1% | 1.8% | 1.8% | 17.1% | 4329 | 736 | 1.8% | 17.0% |
| [0.54, 0.58) | 9665 | 1630 | 8035 | 3.9% | 4.0% | 5.6% | 5.8% | 16.9% | 13994 | 2366 | 5.7% | 16.9% |
| [0.58, 0.63) | 18039 | 3063 | 14976 | 7.3% | 7.4% | 12.9% | 13.2% | 17.0% | 32033 | 5429 | 13.1% | 16.9% |
| [0.63, 0.67) | 55202 | 9528 | 45674 | 22.7% | 22.6% | 35.6% | 35.8% | 17.3% | 87235 | 14957 | 35.8% | 17.1% |
| [0.67, 0.71) | 126327 | 21833 | 104494 | 52.0% | 51.8% | 87.6% | 87.6% | 17.3% | 213562 | 36790 | 87.6% | 17.2% |
| [0.71, 0.76) | 30122 | 5205 | 24917 | 12.4% | 12.4% | 100.0% | 100.0% | 17.3% | 243684 | 41995 | 100.0% | 17.2% |

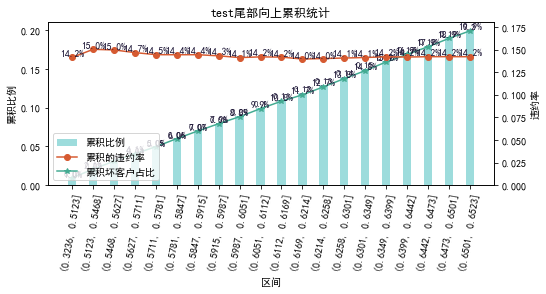


# 八、累积分数分布

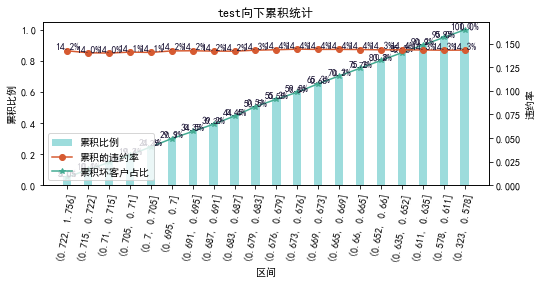
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

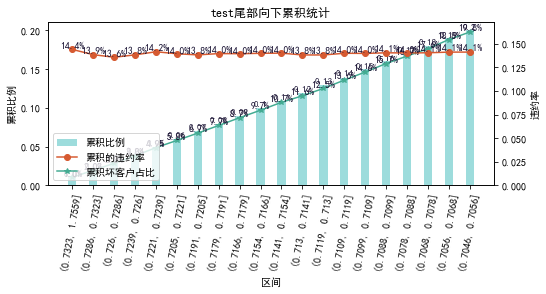
## 1. 向上累积分布





## 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 |
| [0.323, 0.367) | 0.0% | 0.0% | 0.0% | 0.0 | 0.0 |
| [0.367, 0.41) | 0.0% | 0.0% | 0.0% | 0.0 | 0.0 |
| [0.41, 0.453) | 0.2% | 0.2% | 0.2% | 0.0 | 0.0 |
| [0.453, 0.497) | 0.5% | 0.5% | 0.5% | 0.0 | 0.0 |
| [0.497, 0.54) | 1.0% | 1.1% | 1.1% | 0.0 | 0.0 |
| [0.54, 0.583) | 4.0% | 4.0% | 4.0% | 0.0 | 0.0 |
| [0.583, 0.626) | 7.3% | 7.4% | 7.4% | 0.0 | 0.0 |
| [0.626, 0.669) | 22.7% | 22.6% | 22.7% | 0.0 | 0.0 |
| [0.669, 0.713) | 52.0% | 52.1% | 51.8% | 0.0 | 0.0 |
| [0.713, 0.756) | 12.3% | 12.2% | 12.4% | 0.0 | 0.0 |
| sum | 100.0% | 100.0% | 100.0% | 0.0 | 0.0 |

# 十、策略建议