AKULAKU：idpdlold84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 | 2020-10-14 |

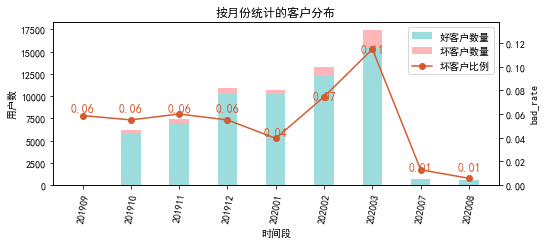
# 一、建模样本

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 |
|  | 67127 | 0.0718 | 201909 | 202002 | 202003 | 202008 |

表2 按月份统计分布表

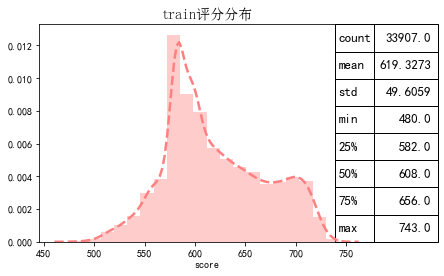
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 201909 | 32 | 2 | 34 | 5.9% | 25.0 | 2.0 | 27.0 | 7.4% | 7.0 | 0.0 | 7.0 | 0.0% |
| 201910 | 5826 | 340 | 6166 | 5.5% | 4122.0 | 237.0 | 4359.0 | 5.4% | 1704.0 | 103.0 | 1807.0 | 5.7% |
| 201911 | 6973 | 446 | 7419 | 6.0% | 4875.0 | 314.0 | 5189.0 | 6.1% | 2098.0 | 132.0 | 2230.0 | 5.9% |
| 201912 | 10294 | 600 | 10894 | 5.5% | 7148.0 | 419.0 | 7567.0 | 5.5% | 3146.0 | 181.0 | 3327.0 | 5.4% |
| 202001 | 10212 | 420 | 10632 | 4.0% | 7138.0 | 297.0 | 7435.0 | 4.0% | 3074.0 | 123.0 | 3197.0 | 3.8% |
| 202002 | 12300 | 994 | 13294 | 7.5% | 8638.0 | 692.0 | 9330.0 | 7.4% | 3662.0 | 302.0 | 3964.0 | 7.6% |
| 202003 | 15456 | 2008 | 17464 | 11.5% | 0.0 | 0.0 | 0.0 | 0.0% | 0.0 | 0.0 | 0.0 | 0.0% |
| 202007 | 687 | 9 | 696 | 1.3% | 0.0 | 0.0 | 0.0 | 0.0% | 0.0 | 0.0 | 0.0 | 0.0% |
| 202008 | 525 | 3 | 528 | 0.6% | 0.0 | 0.0 | 0.0 | 0.0% | 0.0 | 0.0 | 0.0 | 0.0% |

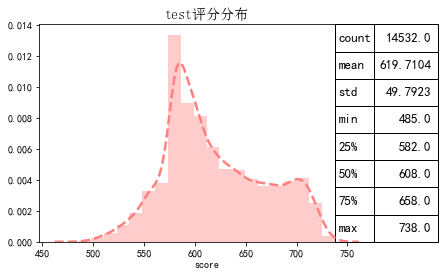


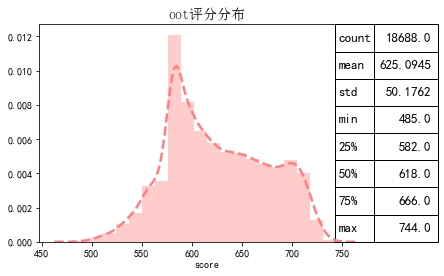
# 二、效果概况

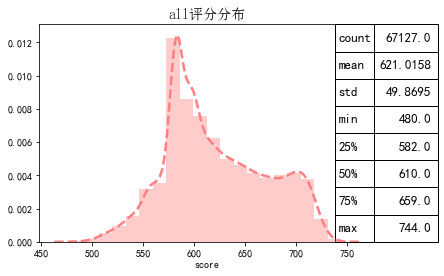
表3 模型效果概况表

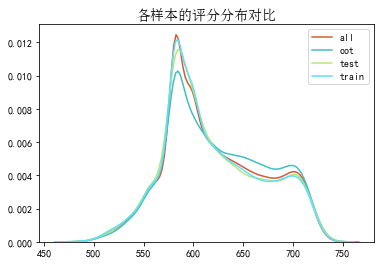
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | train | test | oot | ALL |
| 样本量 | 33907 | 14532 | 18688 | 67127 |
| 坏客户数量 | 1961 | 841 | 2020 | 4822 |
| 坏客户比例 | 5.8% | 5.8% | 10.8% | 7.2% |
| KS | 0.279 | 0.295 | 0.214 | 0.234 |
| AUC | 0.69 | 0.695 | 0.651 | 0.668 |





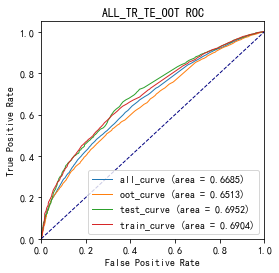
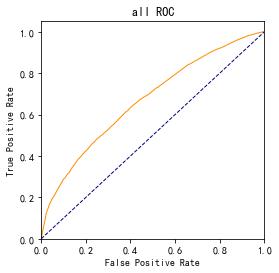
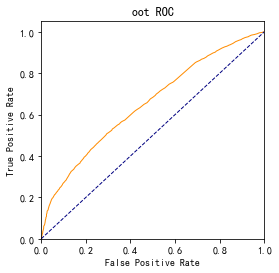
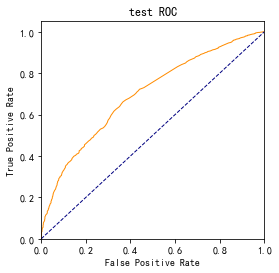
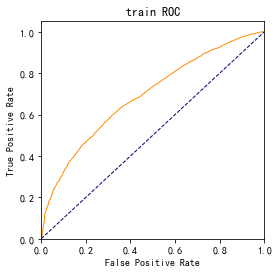






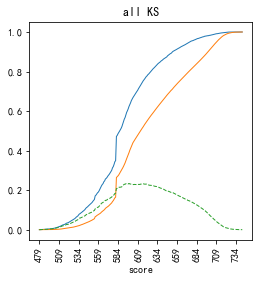
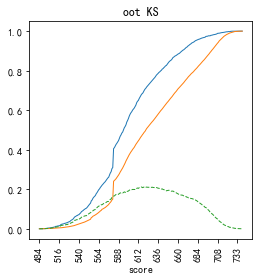
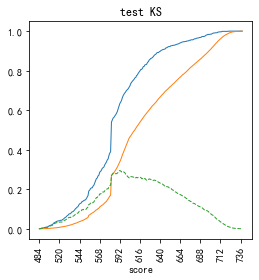
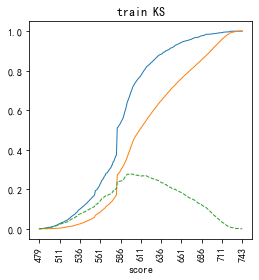
# 三、ROC曲线

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

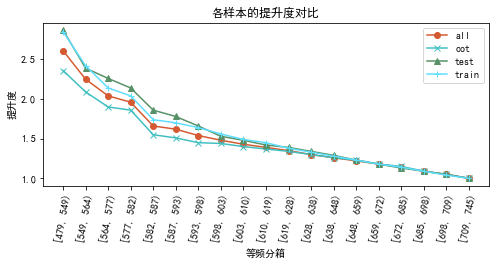
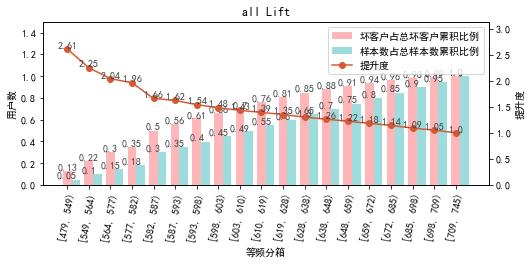
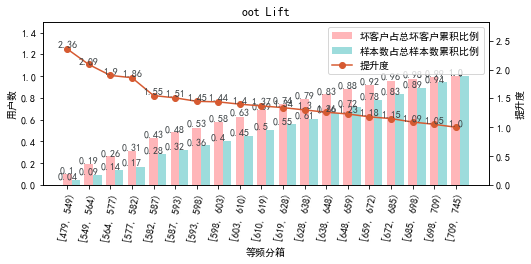
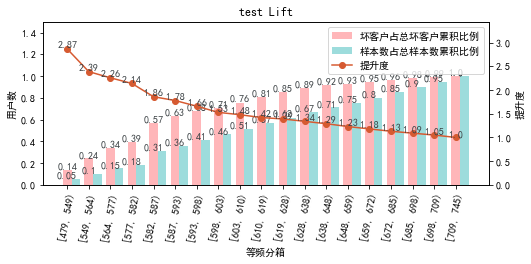
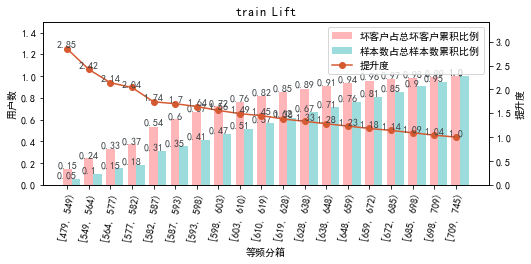
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

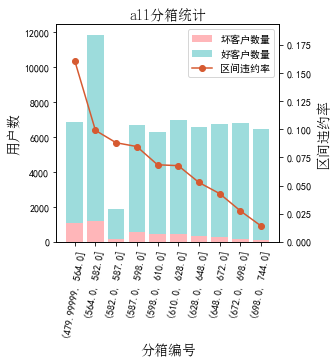
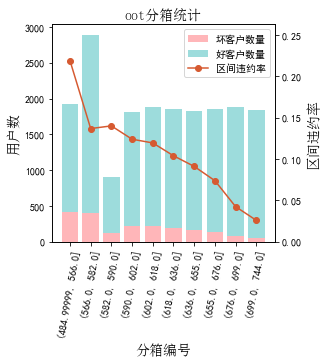
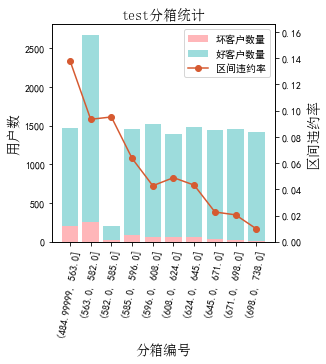
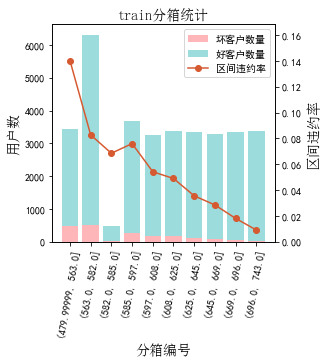
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



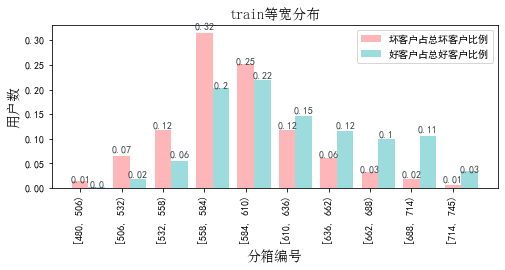
# 七、分数分布

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

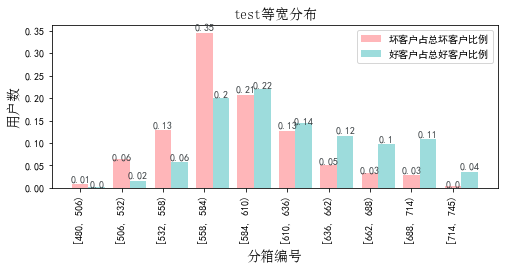
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [480, 506) | 69 | 28 | 41 | 1.4% | 0.1% | 1.4% | 0.1% | 40.6% | 69 | 28 | 0.2% | 40.6% |
| [506, 532) | 687 | 129 | 558 | 6.6% | 1.7% | 8.0% | 1.9% | 18.8% | 756 | 157 | 2.2% | 20.8% |
| [532, 558) | 2002 | 232 | 1770 | 11.8% | 5.5% | 19.8% | 7.4% | 11.6% | 2758 | 389 | 8.1% | 14.1% |
| [558, 584) | 7109 | 619 | 6490 | 31.6% | 20.3% | 51.4% | 27.7% | 8.7% | 9867 | 1008 | 29.1% | 10.2% |
| [584, 610) | 7496 | 492 | 7004 | 25.1% | 21.9% | 76.5% | 49.7% | 6.6% | 17363 | 1500 | 51.2% | 8.6% |
| [610, 636) | 4922 | 230 | 4692 | 11.7% | 14.7% | 88.2% | 64.3% | 4.7% | 22285 | 1730 | 65.7% | 7.8% |
| [636, 662) | 3842 | 120 | 3722 | 6.1% | 11.7% | 94.3% | 76.0% | 3.1% | 26127 | 1850 | 77.1% | 7.1% |
| [662, 688) | 3233 | 65 | 3168 | 3.3% | 9.9% | 97.7% | 85.9% | 2.0% | 29360 | 1915 | 86.6% | 6.5% |
| [688, 714) | 3436 | 36 | 3400 | 1.8% | 10.6% | 99.5% | 96.6% | 1.0% | 32796 | 1951 | 96.7% | 5.9% |
| [714, 745) | 1111 | 10 | 1101 | 0.5% | 3.4% | 100.0% | 100.0% | 0.9% | 33907 | 1961 | 100.0% | 5.8% |



## 2.test

表5 等宽分箱数据分布表

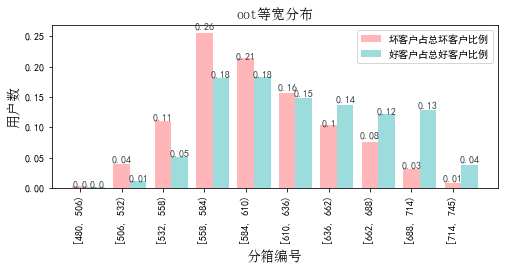
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [480, 506) | 29 | 8 | 21 | 1.0% | 0.2% | 1.0% | 0.2% | 27.6% | 29 | 8 | 0.2% | 27.6% |
| [506, 532) | 262 | 54 | 208 | 6.4% | 1.5% | 7.4% | 1.7% | 20.6% | 291 | 62 | 2.0% | 21.3% |
| [532, 558) | 895 | 109 | 786 | 13.0% | 5.7% | 20.3% | 7.4% | 12.2% | 1186 | 171 | 8.2% | 14.4% |
| [558, 584) | 3029 | 291 | 2738 | 34.6% | 20.0% | 54.9% | 27.4% | 9.6% | 4215 | 462 | 29.0% | 11.0% |
| [584, 610) | 3204 | 174 | 3030 | 20.7% | 22.1% | 75.6% | 49.5% | 5.4% | 7419 | 636 | 51.1% | 8.6% |
| [610, 636) | 2091 | 107 | 1984 | 12.7% | 14.5% | 88.3% | 64.0% | 5.1% | 9510 | 743 | 65.4% | 7.8% |
| [636, 662) | 1624 | 43 | 1581 | 5.1% | 11.5% | 93.5% | 75.6% | 2.6% | 11134 | 786 | 76.6% | 7.1% |
| [662, 688) | 1383 | 28 | 1355 | 3.3% | 9.9% | 96.8% | 85.5% | 2.0% | 12517 | 814 | 86.1% | 6.5% |
| [688, 714) | 1521 | 24 | 1497 | 2.9% | 10.9% | 99.6% | 96.4% | 1.6% | 14038 | 838 | 96.6% | 6.0% |
| [714, 745) | 494 | 3 | 491 | 0.4% | 3.6% | 100.0% | 100.0% | 0.6% | 14532 | 841 | 100.0% | 5.8% |



## 3.oot

表6 等宽分箱数据分布表

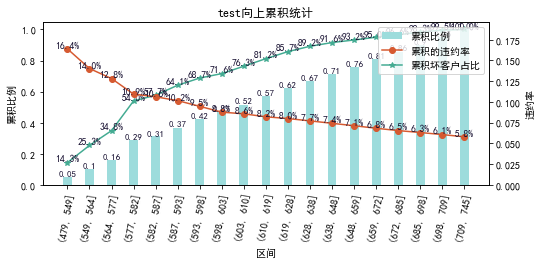
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [480, 506) | 33 | 8 | 25 | 0.4% | 0.1% | 0.4% | 0.1% | 24.2% | 33 | 8 | 0.2% | 24.2% |
| [506, 532) | 274 | 80 | 194 | 4.0% | 1.2% | 4.4% | 1.3% | 29.2% | 307 | 88 | 1.6% | 28.7% |
| [532, 558) | 1061 | 224 | 837 | 11.1% | 5.0% | 15.4% | 6.3% | 21.1% | 1368 | 312 | 7.3% | 22.8% |
| [558, 584) | 3531 | 517 | 3014 | 25.6% | 18.1% | 41.0% | 24.4% | 14.6% | 4899 | 829 | 26.2% | 16.9% |
| [584, 610) | 3475 | 434 | 3041 | 21.5% | 18.2% | 62.5% | 42.7% | 12.5% | 8374 | 1263 | 44.8% | 15.1% |
| [610, 636) | 2790 | 316 | 2474 | 15.6% | 14.8% | 78.2% | 57.5% | 11.3% | 11164 | 1579 | 59.7% | 14.1% |
| [636, 662) | 2483 | 211 | 2272 | 10.4% | 13.6% | 88.6% | 71.1% | 8.5% | 13647 | 1790 | 73.0% | 13.1% |
| [662, 688) | 2172 | 152 | 2020 | 7.5% | 12.1% | 96.1% | 83.3% | 7.0% | 15819 | 1942 | 84.6% | 12.3% |
| [688, 714) | 2216 | 62 | 2154 | 3.1% | 12.9% | 99.2% | 96.2% | 2.8% | 18035 | 2004 | 96.5% | 11.1% |
| [714, 745) | 653 | 16 | 637 | 0.8% | 3.8% | 100.0% | 100.0% | 2.5% | 18688 | 2020 | 100.0% | 10.8% |

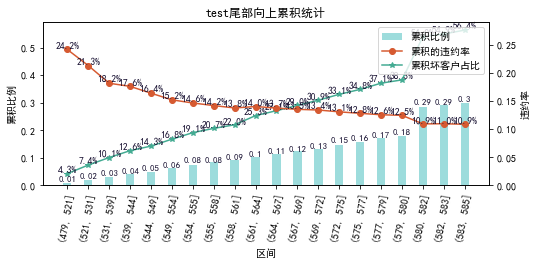


# 八、累积分数分布

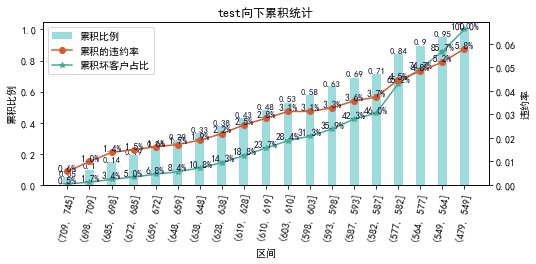
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

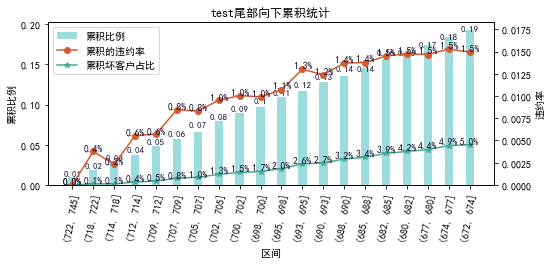
## 1. 向上累积分布





## 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 | oot\_pct | train\_test\_psi | train\_oot\_psi |
| [479.736, 506.4) | 0.2% | 0.3% | 0.2% | 0.0 | 0.0 |
| [506.4, 532.8) | 2.1% | 1.9% | 1.5% | 0.0 | 0.002 |
| [532.8, 559.2) | 6.3% | 6.5% | 6.1% | 0.0 | 0.0 |
| [559.2, 585.6) | 21.4% | 21.4% | 19.4% | 0.0 | 0.002 |
| [585.6, 612.0) | 22.3% | 22.3% | 18.9% | 0.0 | 0.006 |
| [612.0, 638.4) | 14.7% | 14.7% | 15.2% | 0.0 | 0.0 |
| [638.4, 664.8) | 11.2% | 10.7% | 13.1% | 0.0 | 0.003 |
| [664.8, 691.2) | 9.8% | 9.8% | 12.0% | 0.0 | 0.004 |
| [691.2, 717.6) | 10.0% | 10.4% | 11.5% | 0.0 | 0.002 |
| [717.6, 744.0) | 2.0% | 2.1% | 2.1% | 0.0 | 0.0 |
| sum | 100.0% | 100.0% | 100.0% | 0.001 | 0.02 |

# 十、策略建议