AKULAKU：largecashinst4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 | 2020-09-21 |

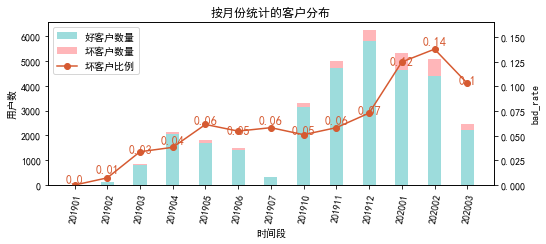
# 一、建模样本

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 |
|  | 34193 | 0.0835 | 201901 | 202002 | 202003 | 202003 |

表2 按月份统计分布表

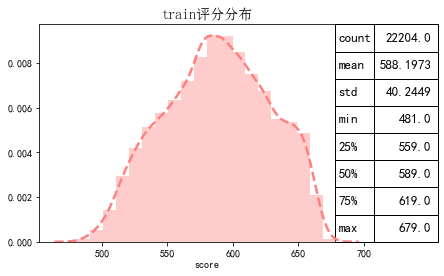
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 201901 | 15 | 0 | 15 | 0.0% | 9.0 | 0.0 | 9.0 | 0.0% | 6.0 | 0.0 | 6.0 | 0.0% |
| 201902 | 136 | 1 | 137 | 0.7% | 95.0 | 1.0 | 96.0 | 1.0% | 41.0 | 0.0 | 41.0 | 0.0% |
| 201903 | 801 | 28 | 829 | 3.4% | 547.0 | 20.0 | 567.0 | 3.5% | 254.0 | 8.0 | 262.0 | 3.1% |
| 201904 | 2064 | 82 | 2146 | 3.8% | 1450.0 | 54.0 | 1504.0 | 3.6% | 614.0 | 28.0 | 642.0 | 4.4% |
| 201905 | 1712 | 112 | 1824 | 6.1% | 1189.0 | 82.0 | 1271.0 | 6.5% | 523.0 | 30.0 | 553.0 | 5.4% |
| 201906 | 1401 | 81 | 1482 | 5.5% | 981.0 | 53.0 | 1034.0 | 5.1% | 420.0 | 28.0 | 448.0 | 6.2% |
| 201907 | 324 | 20 | 344 | 5.8% | 225.0 | 16.0 | 241.0 | 6.6% | 99.0 | 4.0 | 103.0 | 3.9% |
| 201910 | 3143 | 169 | 3312 | 5.1% | 2157.0 | 119.0 | 2276.0 | 5.2% | 986.0 | 50.0 | 1036.0 | 4.8% |
| 201911 | 4712 | 291 | 5003 | 5.8% | 3294.0 | 203.0 | 3497.0 | 5.8% | 1418.0 | 88.0 | 1506.0 | 5.8% |
| 201912 | 5795 | 456 | 6251 | 7.3% | 4034.0 | 316.0 | 4350.0 | 7.3% | 1761.0 | 140.0 | 1901.0 | 7.4% |
| 202001 | 4646 | 662 | 5308 | 12.5% | 3314.0 | 461.0 | 3775.0 | 12.2% | 1332.0 | 201.0 | 1533.0 | 13.1% |
| 202002 | 4371 | 698 | 5069 | 13.8% | 3089.0 | 495.0 | 3584.0 | 13.8% | 1282.0 | 203.0 | 1485.0 | 13.7% |
| 202003 | 2218 | 255 | 2473 | 10.3% | 0.0 | 0.0 | 0.0 | 0.0% | 0.0 | 0.0 | 0.0 | 0.0% |

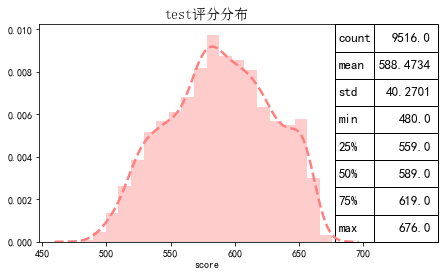


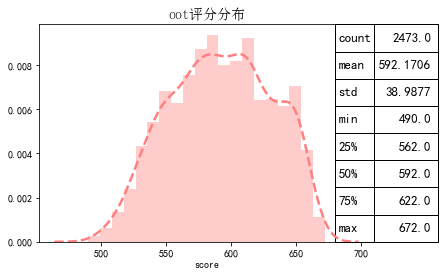
# 二、效果概况

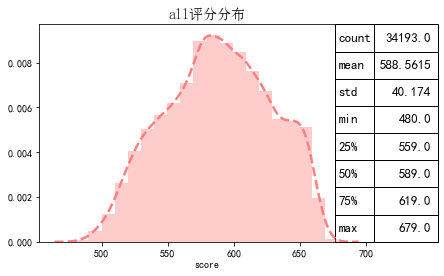
表3 模型效果概况表

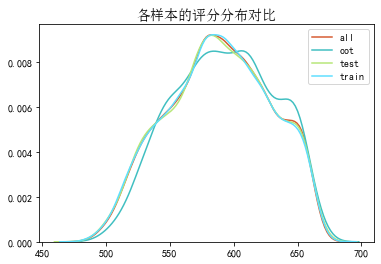
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | train | test | oot | ALL |
| 样本量 | 22204 | 9516 | 2473 | 34193 |
| 坏客户数量 | 1820 | 780 | 255 | 2855 |
| 坏客户比例 | 8.2% | 8.2% | 10.3% | 8.3% |
| KS | 0.265 | 0.253 | 0.232 | 0.255 |
| AUC | 0.679 | 0.68 | 0.658 | 0.677 |





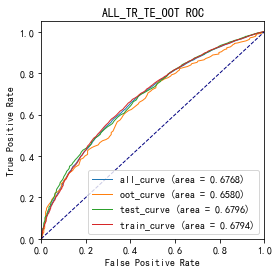
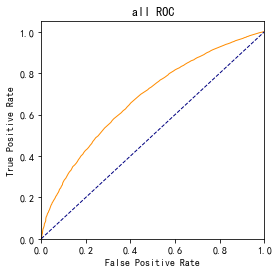
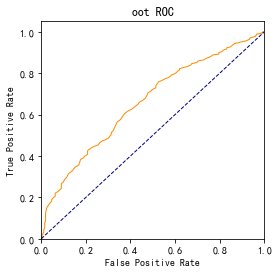
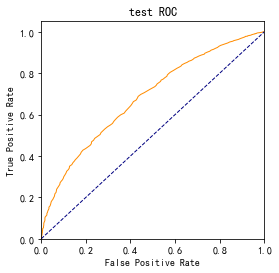
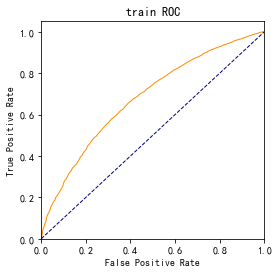






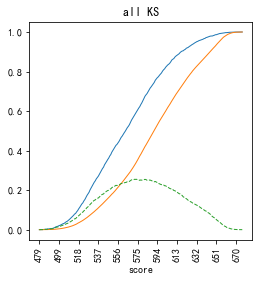
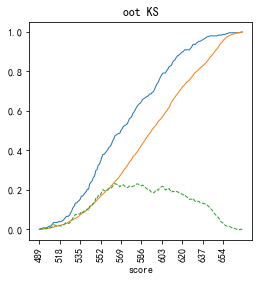
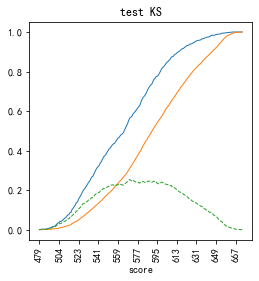
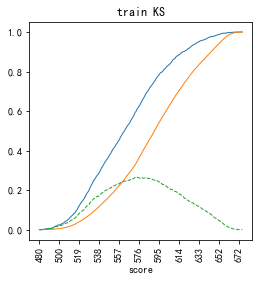
# 三、ROC曲线

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

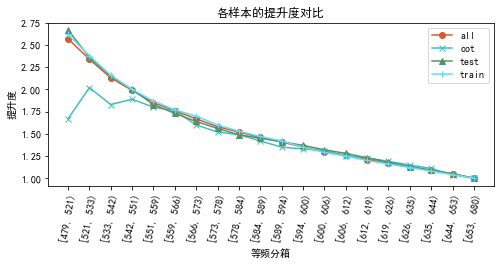
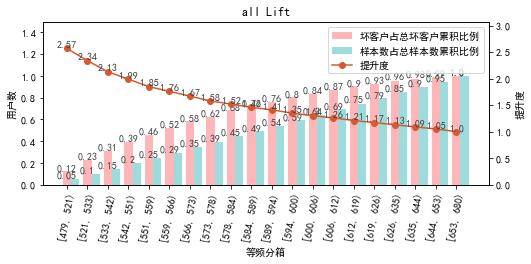
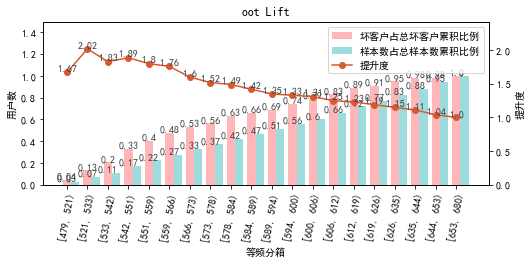
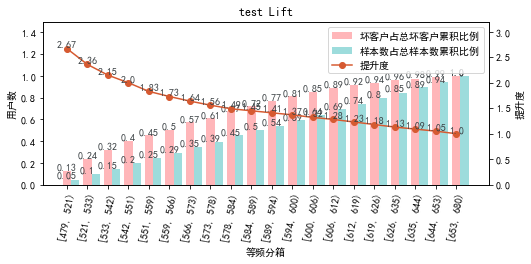
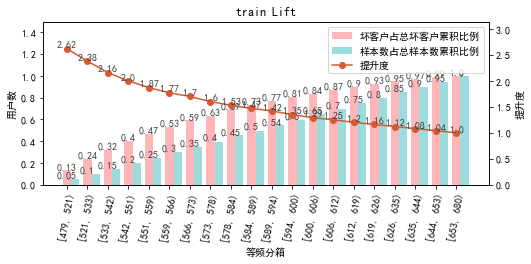
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

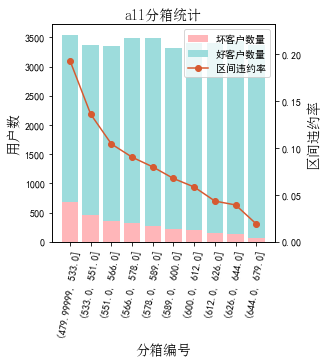
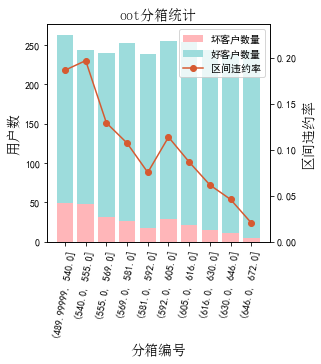
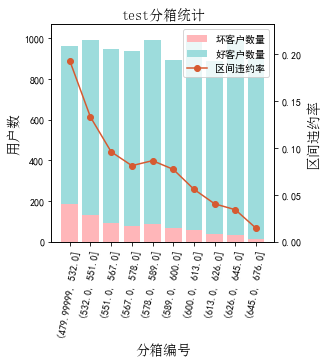
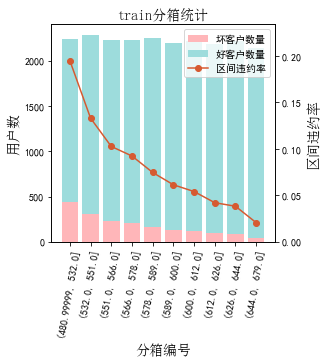
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



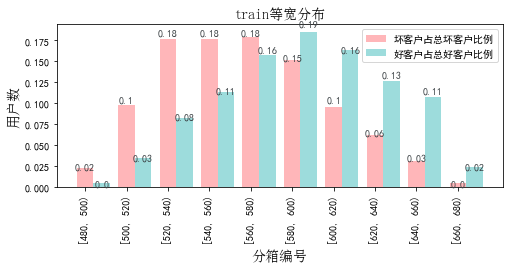
# 七、分数分布

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

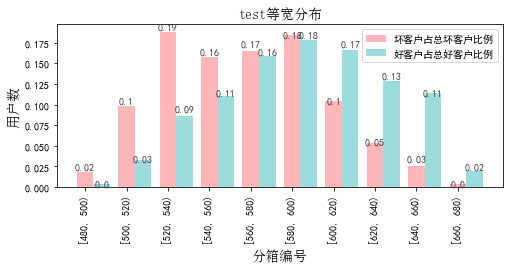
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [480, 500) | 138 | 42 | 96 | 2.3% | 0.5% | 2.3% | 0.5% | 30.4% | 138 | 42 | 0.6% | 30.4% |
| [500, 520) | 890 | 179 | 711 | 9.8% | 3.5% | 12.1% | 4.0% | 20.1% | 1028 | 221 | 4.6% | 21.5% |
| [520, 540) | 2009 | 322 | 1687 | 17.7% | 8.3% | 29.8% | 12.2% | 16.0% | 3037 | 543 | 13.7% | 17.9% |
| [540, 560) | 2625 | 322 | 2303 | 17.7% | 11.3% | 47.5% | 23.5% | 12.3% | 5662 | 865 | 25.5% | 15.3% |
| [560, 580) | 3545 | 327 | 3218 | 18.0% | 15.8% | 65.5% | 39.3% | 9.2% | 9207 | 1192 | 41.5% | 12.9% |
| [580, 600) | 4056 | 276 | 3780 | 15.2% | 18.5% | 80.7% | 57.9% | 6.8% | 13263 | 1468 | 59.7% | 11.1% |
| [600, 620) | 3518 | 173 | 3345 | 9.5% | 16.4% | 90.2% | 74.3% | 4.9% | 16781 | 1641 | 75.6% | 9.8% |
| [620, 640) | 2688 | 114 | 2574 | 6.3% | 12.6% | 96.4% | 86.9% | 4.2% | 19469 | 1755 | 87.7% | 9.0% |
| [640, 660) | 2249 | 57 | 2192 | 3.1% | 10.8% | 99.6% | 97.7% | 2.5% | 21718 | 1812 | 97.8% | 8.3% |
| [660, 680) | 486 | 8 | 478 | 0.4% | 2.3% | 100.0% | 100.0% | 1.6% | 22204 | 1820 | 100.0% | 8.2% |



## 2.test

表5 等宽分箱数据分布表

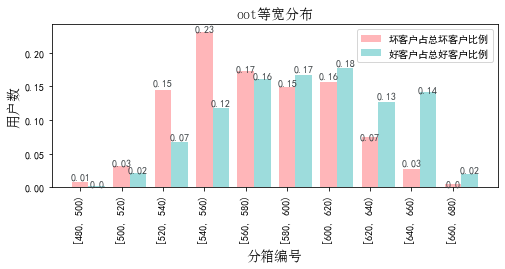
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [480, 500) | 49 | 14 | 35 | 1.8% | 0.4% | 1.8% | 0.4% | 28.6% | 49 | 14 | 0.5% | 28.6% |
| [500, 520) | 368 | 77 | 291 | 9.9% | 3.3% | 11.7% | 3.7% | 20.9% | 417 | 91 | 4.4% | 21.8% |
| [520, 540) | 901 | 147 | 754 | 18.8% | 8.6% | 30.5% | 12.4% | 16.3% | 1318 | 238 | 13.9% | 18.1% |
| [540, 560) | 1085 | 123 | 962 | 15.8% | 11.0% | 46.3% | 23.4% | 11.3% | 2403 | 361 | 25.3% | 15.0% |
| [560, 580) | 1521 | 129 | 1392 | 16.5% | 15.9% | 62.8% | 39.3% | 8.5% | 3924 | 490 | 41.2% | 12.5% |
| [580, 600) | 1705 | 144 | 1561 | 18.5% | 17.9% | 81.3% | 57.2% | 8.4% | 5629 | 634 | 59.2% | 11.3% |
| [600, 620) | 1529 | 81 | 1448 | 10.4% | 16.6% | 91.7% | 73.8% | 5.3% | 7158 | 715 | 75.2% | 10.0% |
| [620, 640) | 1165 | 42 | 1123 | 5.4% | 12.9% | 97.1% | 86.6% | 3.6% | 8323 | 757 | 87.5% | 9.1% |
| [640, 660) | 1017 | 20 | 997 | 2.6% | 11.4% | 99.6% | 98.0% | 2.0% | 9340 | 777 | 98.2% | 8.3% |
| [660, 680) | 176 | 3 | 173 | 0.4% | 2.0% | 100.0% | 100.0% | 1.7% | 9516 | 780 | 100.0% | 8.2% |



## 3.oot

表6 等宽分箱数据分布表

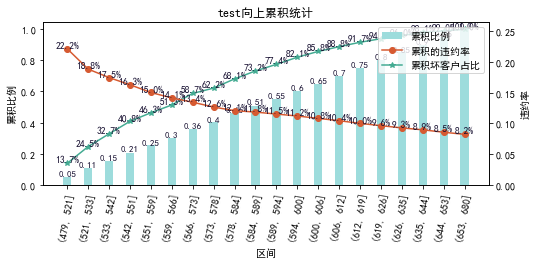
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [480, 500) | 6 | 2 | 4 | 0.8% | 0.2% | 0.8% | 0.2% | 33.3% | 6 | 2 | 0.2% | 33.3% |
| [500, 520) | 53 | 8 | 45 | 3.1% | 2.0% | 3.9% | 2.2% | 15.1% | 59 | 10 | 2.4% | 16.9% |
| [520, 540) | 187 | 37 | 150 | 14.5% | 6.8% | 18.4% | 9.0% | 19.8% | 246 | 47 | 9.9% | 19.1% |
| [540, 560) | 321 | 59 | 262 | 23.1% | 11.8% | 41.6% | 20.8% | 18.4% | 567 | 106 | 22.9% | 18.7% |
| [560, 580) | 401 | 44 | 357 | 17.3% | 16.1% | 58.8% | 36.9% | 11.0% | 968 | 150 | 39.1% | 15.5% |
| [580, 600) | 408 | 38 | 370 | 14.9% | 16.7% | 73.7% | 53.6% | 9.3% | 1376 | 188 | 55.6% | 13.7% |
| [600, 620) | 432 | 40 | 392 | 15.7% | 17.7% | 89.4% | 71.2% | 9.3% | 1808 | 228 | 73.1% | 12.6% |
| [620, 640) | 300 | 19 | 281 | 7.5% | 12.7% | 96.9% | 83.9% | 6.3% | 2108 | 247 | 85.2% | 11.7% |
| [640, 660) | 320 | 7 | 313 | 2.7% | 14.1% | 99.6% | 98.0% | 2.2% | 2428 | 254 | 98.2% | 10.5% |
| [660, 680) | 45 | 1 | 44 | 0.4% | 2.0% | 100.0% | 100.0% | 2.2% | 2473 | 255 | 100.0% | 10.3% |

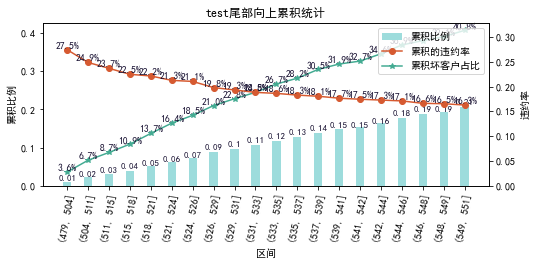


# 八、累积分数分布

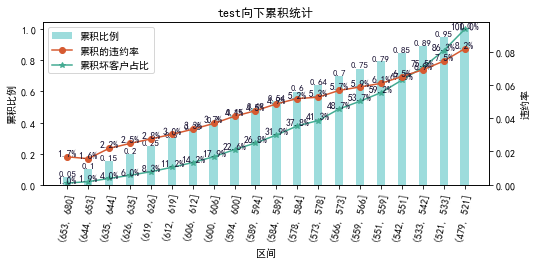
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

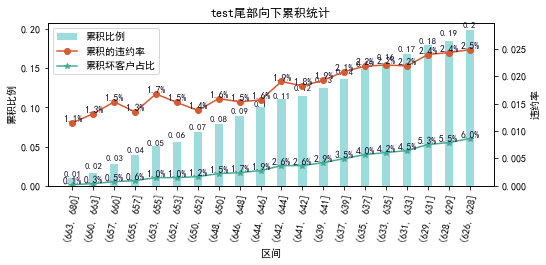
## 1. 向上累积分布





## 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 |
| [479.801, 499.9) | 0.6% | 0.5% | 0.2% | 0.0 | 0.004 |
| [499.9, 519.8) | 4.0% | 3.9% | 2.1% | 0.0 | 0.012 |
| [519.8, 539.7) | 9.0% | 9.5% | 7.6% | 0.0 | 0.003 |
| [539.7, 559.6) | 11.8% | 11.4% | 13.0% | 0.0 | 0.001 |
| [559.6, 579.5) | 16.0% | 16.0% | 16.2% | 0.0 | 0.0 |
| [579.5, 599.4) | 18.3% | 17.9% | 16.5% | 0.0 | 0.002 |
| [599.4, 619.3) | 15.8% | 16.1% | 17.5% | 0.0 | 0.002 |
| [619.3, 639.2) | 12.1% | 12.2% | 12.1% | 0.0 | 0.0 |
| [639.2, 659.1) | 10.1% | 10.7% | 12.9% | 0.0 | 0.007 |
| [659.1, 679.0) | 2.2% | 1.8% | 1.8% | 0.001 | 0.001 |
| sum | 100.0% | 100.0% | 100.0% | 0.002 | 0.03 |

# 十、策略建议