AKULAKU：largecashinst6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 |

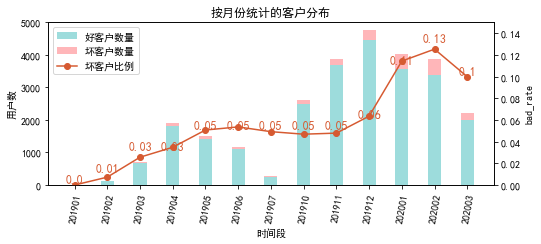
# 一、建模样本

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 |
|  | 26999 | 0.0745 | 201901 | 202002 | 202003 | 202003 |

表2 按月份统计分布表

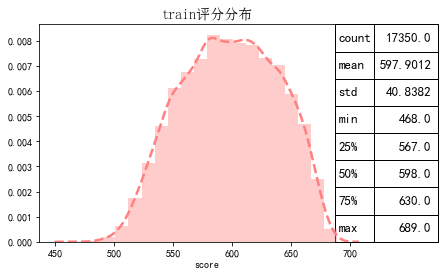
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 好客户数 | 坏客户数 | 总数 | 总违约率 | 训练集好客户数 | 训练集坏客户数 | 训练集总数 | 训练集违约率 | 测试集好客户数 | 测试集坏客户数 | 测试集总数 | 测试集违约率 |
| 201901 | 15 | 0 | 15 | 0.0% | 10.0 | 0.0 | 10.0 | 0.0% | 5.0 | 0.0 | 5.0 | 0.0% |
| 201902 | 136 | 1 | 137 | 0.7% | 89.0 | 1.0 | 90.0 | 1.1% | 47.0 | 0.0 | 47.0 | 0.0% |
| 201903 | 683 | 18 | 701 | 2.6% | 493.0 | 11.0 | 504.0 | 2.2% | 190.0 | 7.0 | 197.0 | 3.6% |
| 201904 | 1828 | 66 | 1894 | 3.5% | 1260.0 | 44.0 | 1304.0 | 3.4% | 568.0 | 22.0 | 590.0 | 3.7% |
| 201905 | 1421 | 76 | 1497 | 5.1% | 1014.0 | 51.0 | 1065.0 | 4.8% | 407.0 | 25.0 | 432.0 | 5.8% |
| 201906 | 1094 | 62 | 1156 | 5.4% | 752.0 | 41.0 | 793.0 | 5.2% | 342.0 | 21.0 | 363.0 | 5.8% |
| 201907 | 251 | 13 | 264 | 4.9% | 155.0 | 9.0 | 164.0 | 5.5% | 96.0 | 4.0 | 100.0 | 4.0% |
| 201910 | 2478 | 122 | 2600 | 4.7% | 1748.0 | 77.0 | 1825.0 | 4.2% | 730.0 | 45.0 | 775.0 | 5.8% |
| 201911 | 3677 | 185 | 3862 | 4.8% | 2598.0 | 126.0 | 2724.0 | 4.6% | 1079.0 | 59.0 | 1138.0 | 5.2% |
| 201912 | 4466 | 303 | 4769 | 6.4% | 3132.0 | 217.0 | 3349.0 | 6.5% | 1334.0 | 86.0 | 1420.0 | 6.1% |
| 202001 | 3563 | 460 | 4023 | 11.4% | 2486.0 | 329.0 | 2815.0 | 11.7% | 1077.0 | 131.0 | 1208.0 | 10.8% |
| 202002 | 3384 | 485 | 3869 | 12.5% | 2359.0 | 348.0 | 2707.0 | 12.9% | 1025.0 | 137.0 | 1162.0 | 11.8% |
| 202003 | 1991 | 221 | 2212 | 10.0% | 0.0 | 0.0 | 0.0 | 0.0% | 0.0 | 0.0 | 0.0 | 0.0% |

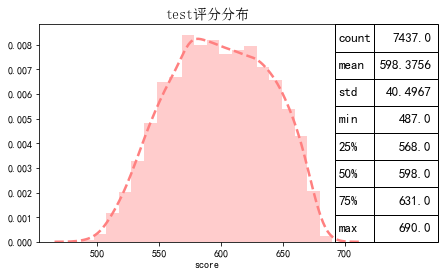


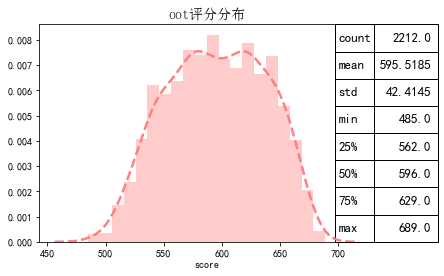
# 二、效果概况

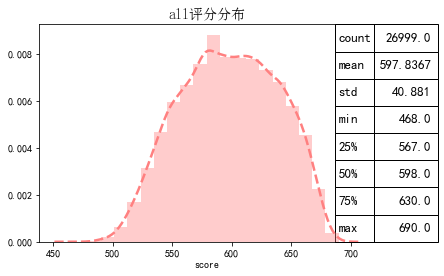
表3 模型效果概况表

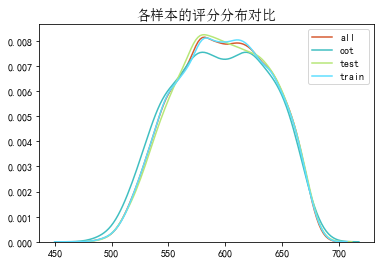
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | train | test | oot | ALL |
| 样本量 | 17350 | 7437 | 2212 | 26999 |
| 坏客户数量 | 1254 | 537 | 221 | 2012 |
| 坏客户比例 | 7.2% | 7.2% | 10.0% | 7.5% |
| KS | 0.293 | 0.294 | 0.283 | 0.289 |
| AUC | 0.693 | 0.698 | 0.686 | 0.694 |





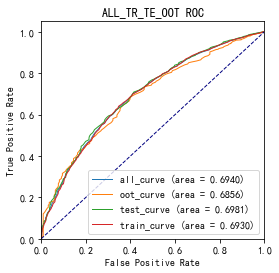
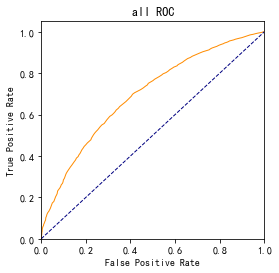
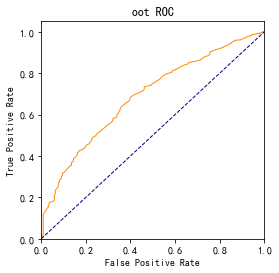
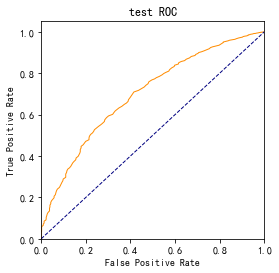
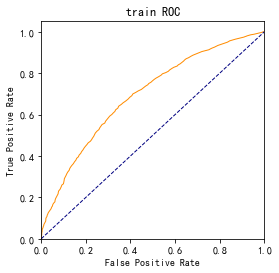






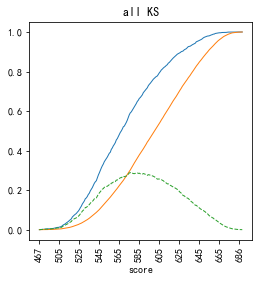
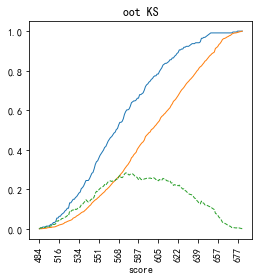
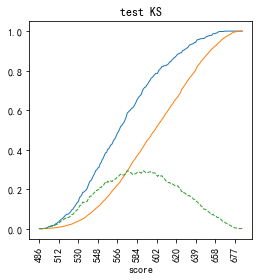
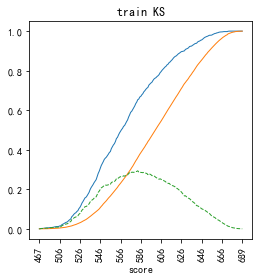
# 三、ROC曲线

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

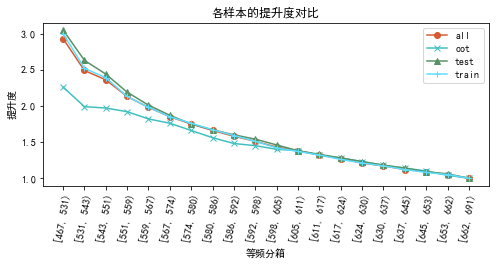
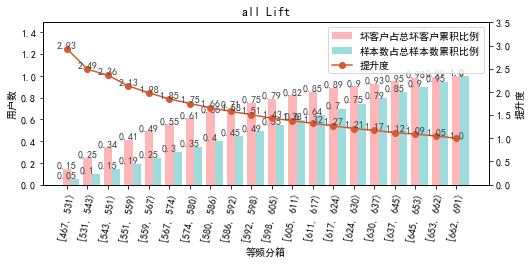
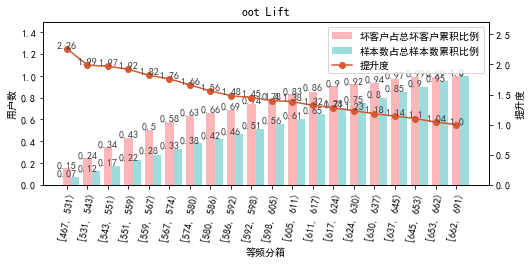
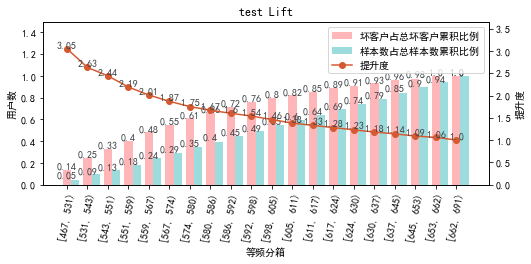
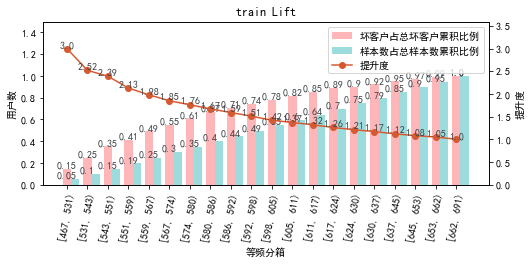
KS（Kolmogorov-Smirnov）：通过衡量好坏样本累计分布之间的差值，来评估模型的风险区分能力。KS值越大，区分度越强。



# 五、Lift曲线

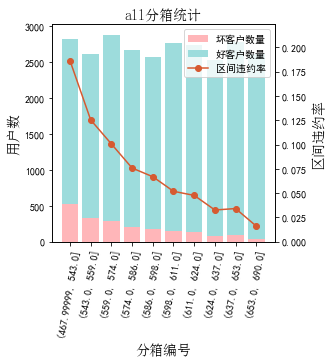
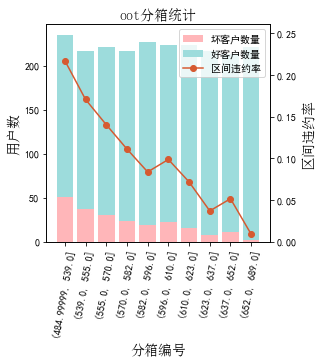
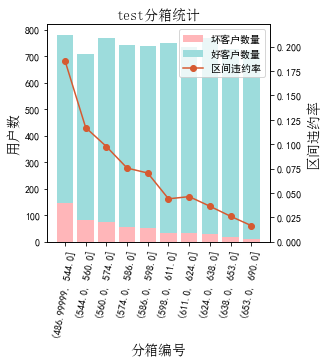
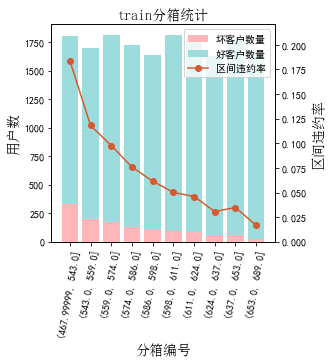
Lift：提升度，表示与不利用模型相比，模型的预测能力提升多少，lift(提升指数)越大，模型效果越好。

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。



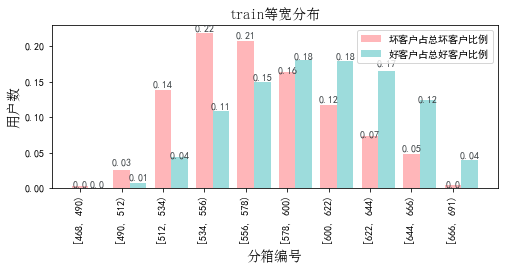
# 七、分数分布

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

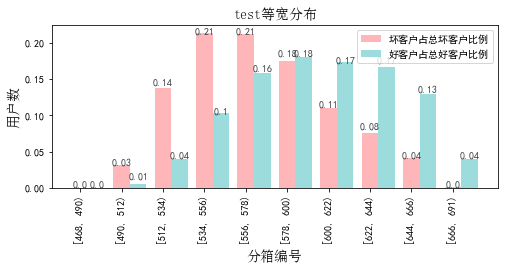
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [468, 490) | 7 | 4 | 3 | 0.3% | 0.0% | 0.3% | 0.0% | 57.1% | 7 | 4 | 0.0% | 57.1% |
| [490, 512) | 142 | 32 | 110 | 2.6% | 0.7% | 2.9% | 0.7% | 22.5% | 149 | 36 | 0.9% | 24.2% |
| [512, 534) | 880 | 173 | 707 | 13.8% | 4.4% | 16.7% | 5.1% | 19.7% | 1029 | 209 | 5.9% | 20.3% |
| [534, 556) | 2031 | 275 | 1756 | 21.9% | 10.9% | 38.6% | 16.0% | 13.5% | 3060 | 484 | 17.6% | 15.8% |
| [556, 578) | 2660 | 261 | 2399 | 20.8% | 14.9% | 59.4% | 30.9% | 9.8% | 5720 | 745 | 33.0% | 13.0% |
| [578, 600) | 3109 | 205 | 2904 | 16.3% | 18.0% | 75.8% | 49.0% | 6.6% | 8829 | 950 | 50.9% | 10.8% |
| [600, 622) | 3040 | 147 | 2893 | 11.7% | 18.0% | 87.5% | 66.9% | 4.8% | 11869 | 1097 | 68.4% | 9.2% |
| [622, 644) | 2761 | 92 | 2669 | 7.3% | 16.6% | 94.8% | 83.5% | 3.3% | 14630 | 1189 | 84.3% | 8.1% |
| [644, 666) | 2069 | 60 | 2009 | 4.8% | 12.5% | 99.6% | 96.0% | 2.9% | 16699 | 1249 | 96.2% | 7.5% |
| [666, 691) | 651 | 5 | 646 | 0.4% | 4.0% | 100.0% | 100.0% | 0.8% | 17350 | 1254 | 100.0% | 7.2% |



## 2.test

表5 等宽分箱数据分布表

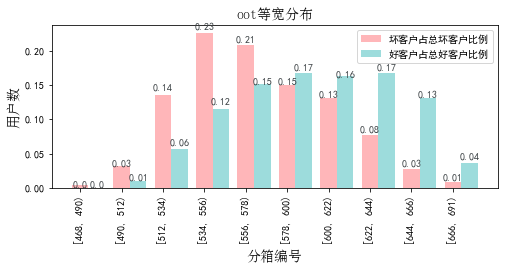
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [468, 490) | 2 | 0 | 2 | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 2 | 0 | 0.0% | 0.0% |
| [490, 512) | 59 | 17 | 42 | 3.2% | 0.6% | 3.2% | 0.6% | 28.8% | 61 | 17 | 0.8% | 27.9% |
| [512, 534) | 348 | 74 | 274 | 13.8% | 4.0% | 16.9% | 4.6% | 21.3% | 409 | 91 | 5.5% | 22.2% |
| [534, 556) | 828 | 115 | 713 | 21.4% | 10.3% | 38.4% | 14.9% | 13.9% | 1237 | 206 | 16.6% | 16.7% |
| [556, 578) | 1212 | 114 | 1098 | 21.2% | 15.9% | 59.6% | 30.9% | 9.4% | 2449 | 320 | 32.9% | 13.1% |
| [578, 600) | 1342 | 94 | 1248 | 17.5% | 18.1% | 77.1% | 48.9% | 7.0% | 3791 | 414 | 51.0% | 10.9% |
| [600, 622) | 1260 | 59 | 1201 | 11.0% | 17.4% | 88.1% | 66.3% | 4.7% | 5051 | 473 | 67.9% | 9.4% |
| [622, 644) | 1195 | 41 | 1154 | 7.6% | 16.7% | 95.7% | 83.1% | 3.4% | 6246 | 514 | 84.0% | 8.2% |
| [644, 666) | 917 | 22 | 895 | 4.1% | 13.0% | 99.8% | 96.0% | 2.4% | 7163 | 536 | 96.3% | 7.5% |
| [666, 691) | 274 | 1 | 273 | 0.2% | 4.0% | 100.0% | 100.0% | 0.4% | 7437 | 537 | 100.0% | 7.2% |



## 3.oot

表6 等宽分箱数据分布表

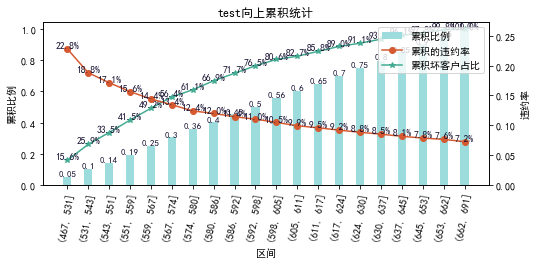
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 组内总人数 | 组内坏客户数 | 组内好客户数 | 坏客户占比 | 好客户占比 | 累计坏客户占比 | 累计好客户占比 | 区间违约率 | 累计拒绝人数 | 累计拒绝坏人数 | 累计拒绝率 | 累计拒绝坏人占比 |
| [468, 490) | 2 | 1 | 1 | 0.5% | 0.1% | 0.5% | 0.1% | 50.0% | 2 | 1 | 0.1% | 50.0% |
| [490, 512) | 28 | 7 | 21 | 3.2% | 1.1% | 3.6% | 1.1% | 25.0% | 30 | 8 | 1.4% | 26.7% |
| [512, 534) | 143 | 30 | 113 | 13.6% | 5.7% | 17.2% | 6.8% | 21.0% | 173 | 38 | 7.8% | 22.0% |
| [534, 556) | 279 | 50 | 229 | 22.6% | 11.5% | 39.8% | 18.3% | 17.9% | 452 | 88 | 20.4% | 19.5% |
| [556, 578) | 348 | 46 | 302 | 20.8% | 15.2% | 60.6% | 33.5% | 13.2% | 800 | 134 | 36.2% | 16.8% |
| [578, 600) | 365 | 33 | 332 | 14.9% | 16.7% | 75.6% | 50.1% | 9.0% | 1165 | 167 | 52.7% | 14.3% |
| [600, 622) | 355 | 29 | 326 | 13.1% | 16.4% | 88.7% | 66.5% | 8.2% | 1520 | 196 | 68.7% | 12.9% |
| [622, 644) | 349 | 17 | 332 | 7.7% | 16.7% | 96.4% | 83.2% | 4.9% | 1869 | 213 | 84.5% | 11.4% |
| [644, 666) | 268 | 6 | 262 | 2.7% | 13.2% | 99.1% | 96.3% | 2.2% | 2137 | 219 | 96.6% | 10.2% |
| [666, 691) | 75 | 2 | 73 | 0.9% | 3.7% | 100.0% | 100.0% | 2.7% | 2212 | 221 | 100.0% | 10.0% |

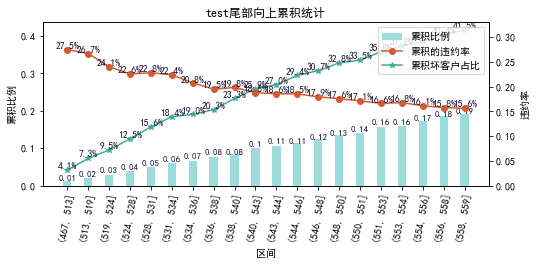


# 八、累积分数分布

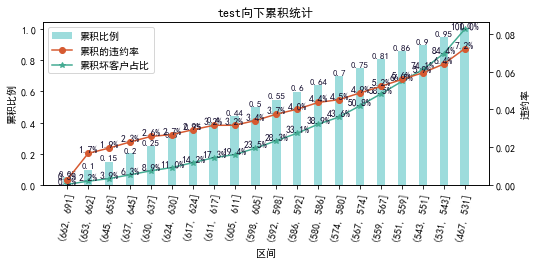
测试集等频分为20个区间，进行向上累积、向下累积。可通过向上累积选择拒绝阈值，通过向下累积选择通过阈值。

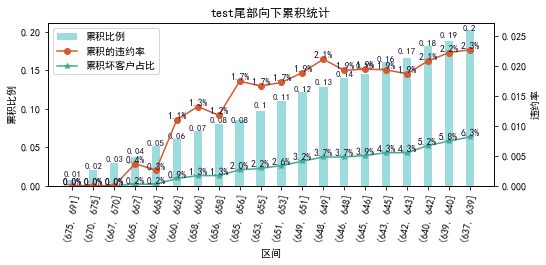
## 1. 向上累积分布





## 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 |
| [467.778, 490.2) | 0.0% | 0.0% | 0.2% | 0.0 | 0.002 |
| [490.2, 512.4) | 0.9% | 0.9% | 1.4% | 0.0 | 0.002 |
| [512.4, 534.6) | 5.4% | 5.1% | 6.7% | 0.0 | 0.003 |
| [534.6, 556.8) | 11.9% | 11.3% | 12.7% | 0.0 | 0.001 |
| [556.8, 579.0) | 15.5% | 16.6% | 16.1% | 0.001 | 0.0 |
| [579.0, 601.2) | 18.7% | 18.7% | 16.5% | 0.0 | 0.003 |
| [601.2, 623.4) | 17.6% | 17.0% | 17.0% | 0.0 | 0.0 |
| [623.4, 645.6) | 15.6% | 16.0% | 15.4% | 0.0 | 0.0 |
| [645.6, 667.8) | 11.5% | 11.6% | 11.3% | 0.0 | 0.0 |
| [667.8, 690.0) | 2.9% | 2.9% | 2.8% | 0.0 | 0.0 |
| sum | 100.0% | 100.0% | 100.0% | 0.002 | 0.011 |

# 十、策略建议