**针对金融客群**

**基于深度学习模型的**

**POC模型报告**

**大连福瑞泽科技有限公司**

**2018年06月**

目录

[模型训练 3](#_Toc529699410)

[**1.** **数据预处理** 3](#_Toc529699411)

[**2.** **卷积神经网络模型详述** 5](#_Toc529699412)

[**3.** **训练过程** 6](#_Toc529699413)

[**4.** **测试过程** 7](#_Toc529699414)

[一、 模型基本指标 7](#_Toc529699415)

[二、 模型KS与AUC图像 8](#_Toc529699416)

[三、 模型输出分布对比及PSI 13](#_Toc529699417)

[四、 分布宽表 14](#_Toc529699418)

# 模型训练

## **数据预处理**

训练数据的处理分为两个部分：第一部分为特征处理，包含缺失值填充、特征增补和one-hot编码；第二部分是自编码机压缩。最后再将两部分的结果进行拼接。本阶段的处理过程如图1所示。

1. 特征处理

缺失值填充：我们分别用-1和-2替换样本中的缺失值-8888和-9999，这样可以使数据更偏向中心。

One-hot编码：one-hot编码是指使用N位状态寄存器来对N个状态进行编码，每个状态都有独立的寄存器位，在任意样本的one-hot编码中只有一位有效。例如，如果两个样本的某个原始特征取值分别为0和1，经过编码后分别为10和01。

原始特征为270列，我们丢弃了缺失率超过70%的属性，总数为39列；对缺失率在30%-70%的属性做了One-hot编码，这一部分最终得到了80列；对缺失率在30%以下的属性仅做了缺失值填充，这一部分总共占151列。经过上述特征处理最终得到270-39+80=311列特征。

1. 自编码机压缩

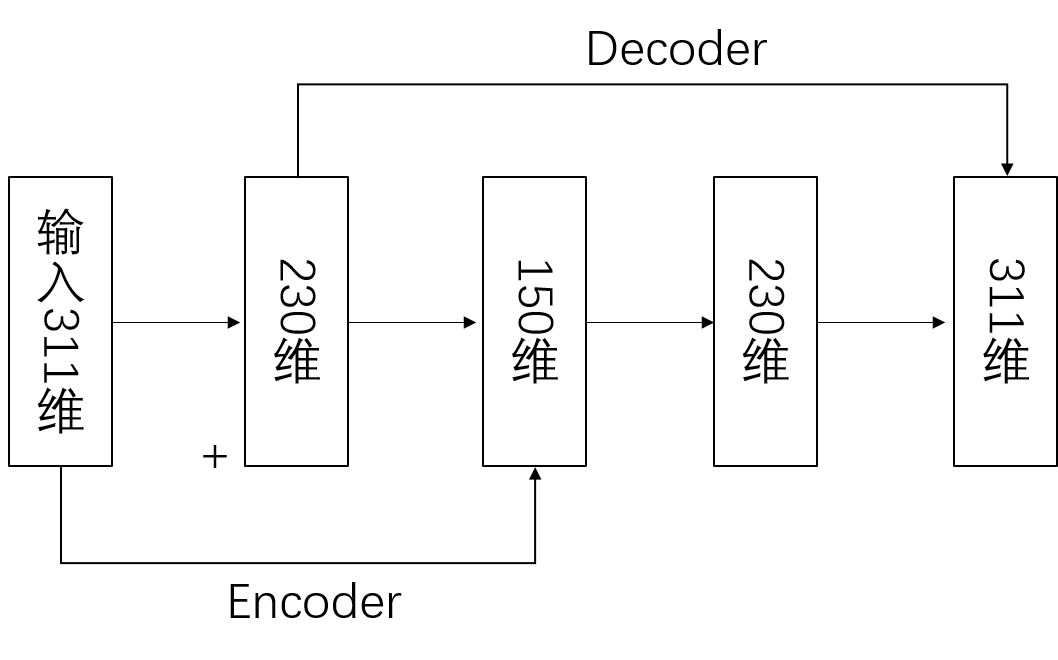
自编码机是一种无监督的数据压缩算法，可以看做一种数据降维算法，采用自编码算法能有效地将离散化数据转化成连续的数据，可以为训练带来帮助。我们的自编码模型是采用BP神经网络结构，如图2所示。其中Dense层的激活函

图1：AutorEncoder结构

数分别为 relu，线性激活,relu, sigmoid最后一层为了增强表达能力采取了sigmoid激活函数，取中间层的150维为输出。

这一部分使用自编码机对第一部分得到的311维特征进行自编码处理，经过自编码训练后得到150维的压缩后的特征。

1. 特征拼接

将特征处理得到的311维特征与自编码机压缩得到的150维特征进行拼接，一起组成最终的461维训练特征，用于下一步神经网络的输入。

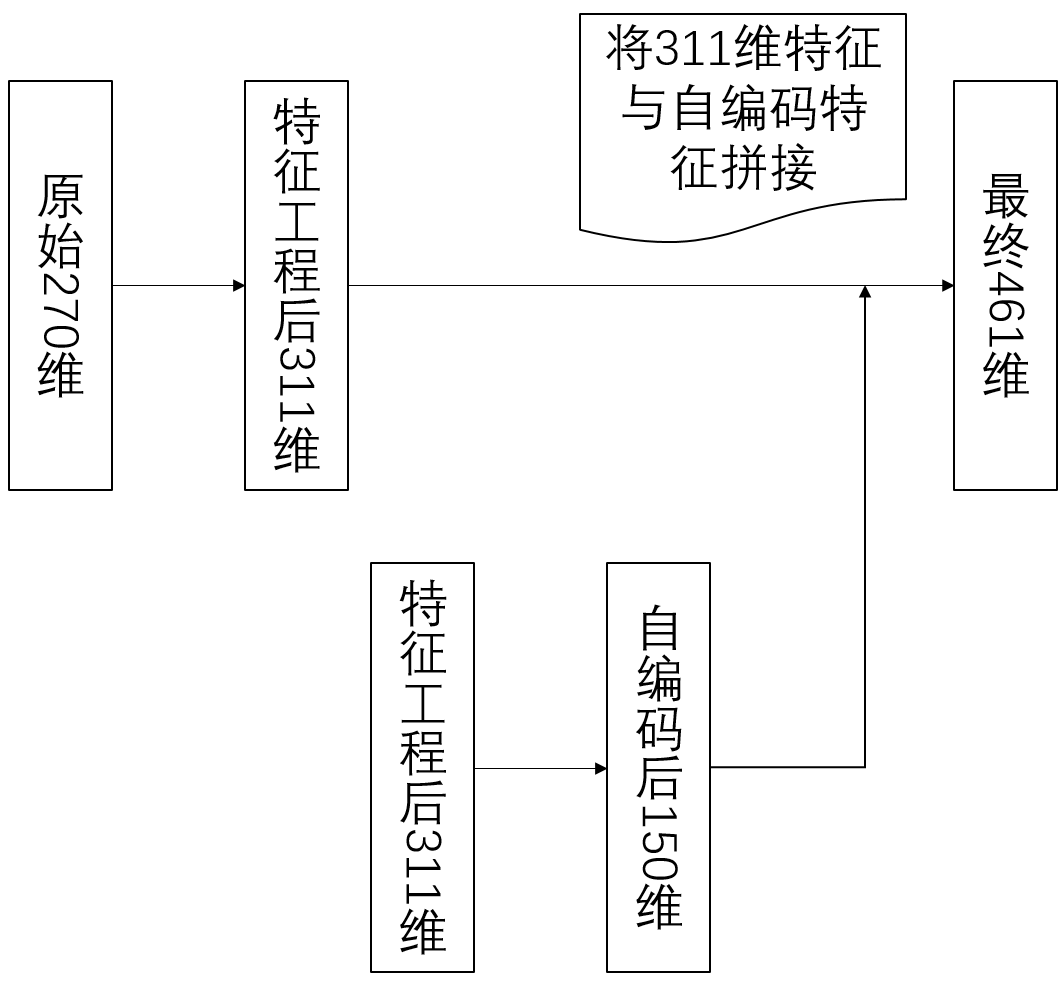


图 2：数据预处理阶段特征整体过程示意图

## **卷积神经网络模型详述**

我们使用的网络采取类似于Alexnet结构，包含8层，输入为数据预处理部分得到的463维特征，输出为样本违约的概率。由于训练训练数据非图像数据，无法直接进行卷积，因此我们将数据升高一维，变成3维结构，在训练时候就可以正常进行卷积处理。我们的神经网络模型示意图如下图3所示：

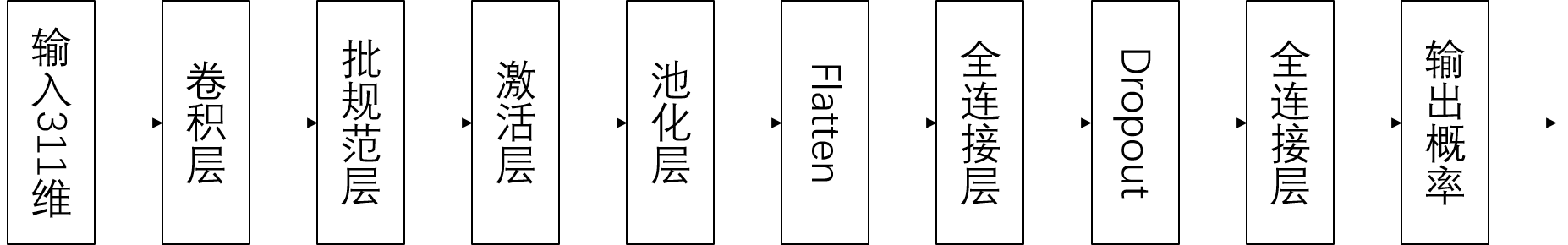


图3：网络结构

网络各层参数说明表格：

|  |  |
| --- | --- |
| 网络层 | 参数 |
| Conv1D | Channels = 32, kernel\_size =33, padding = 16 |
| BatchNorm | axis = 1(对通道进行规范化) |
| Activation | Activation=‘relu’ |
| MaxPool1D | pool\_size=13 |
| Flatten |  |
| Dense | Out\_dim = 33, activation= ‘relu’ |
| Dense | Out\_dim = 1, activation = ‘sigmoid’ |
| Dropout | 0.2 |
| Output | Probability of bad person |

## 表1：网络结构

## **训练过程**

1. 欠采样

考虑到该金融数据的不平衡性（好人与坏人的比例接近100:1），因此采取必要的不平衡的抽样方法是必不可或缺的。经过调查研究，我们这里采取了BalanceCascade欠采样方法，即在神经网络的每一轮训练时都使用多数类与少数类数量上相等的一个batch，以保证模型的平衡性。

1. 十折交叉验证

经过数据预处理得到的样本共有311维特征，对数据集进行10折交叉验证，即把数据集随机分为10份，每次取8份作为train，1份作为val，1份作为test，进行使用，共进行10次，保证每份都做一次val和test。取10次试验分别得到的AUC与KS的均值作为超参数选择的标准，训练完毕后保存模型。

## **测试过程**

1. 预测结果

训练完深度模型后，输入所有的训练数据，可以得到训练数据的预测为好人/坏人的概率，考虑到数据的不平衡性，训练数据的好人预测结果绝大部分概率值右倾（分类器倾向于所有人都预测为好人）；另一方面，只要不改变原来预测概率值的排序顺序，模型的ks，auc值都不会改变，因此为了模型的稳定性，我们生成和训练样本数量一致的，概率质量分布呈正态分布的点，并将训练样本的概率值，按照从小到大的顺序一一映射到该曲线上。对于在test阶段，我们首先将数据经过预处理生成311维的新特征，再和原始的150维特征一起拼接成461维特征，经过神经网络后，得到最初的预测概率值，然后通过knn近邻的思想，找到训练数据中最靠近该概率值的点，并将该点映射到正态分布曲线上的点的值作为最终的预测概率，通过这样的后处理能够使得模型的稳定性大大增加。

# 模型基本指标

根据福瑞泽输出评分，对模型在三个数据集上的基本指标进行计算，计算时使用返回数据中good\_prob字段作为输出分数，结果如下：

|  |  |  |
| --- | --- | --- |
| 数据集 | KS | AUC |
| TRAIN\_06\_12 | 0.4710 | 0.8116 |
| TEST\_01\_02 | 0.4153 | 0.7694 |

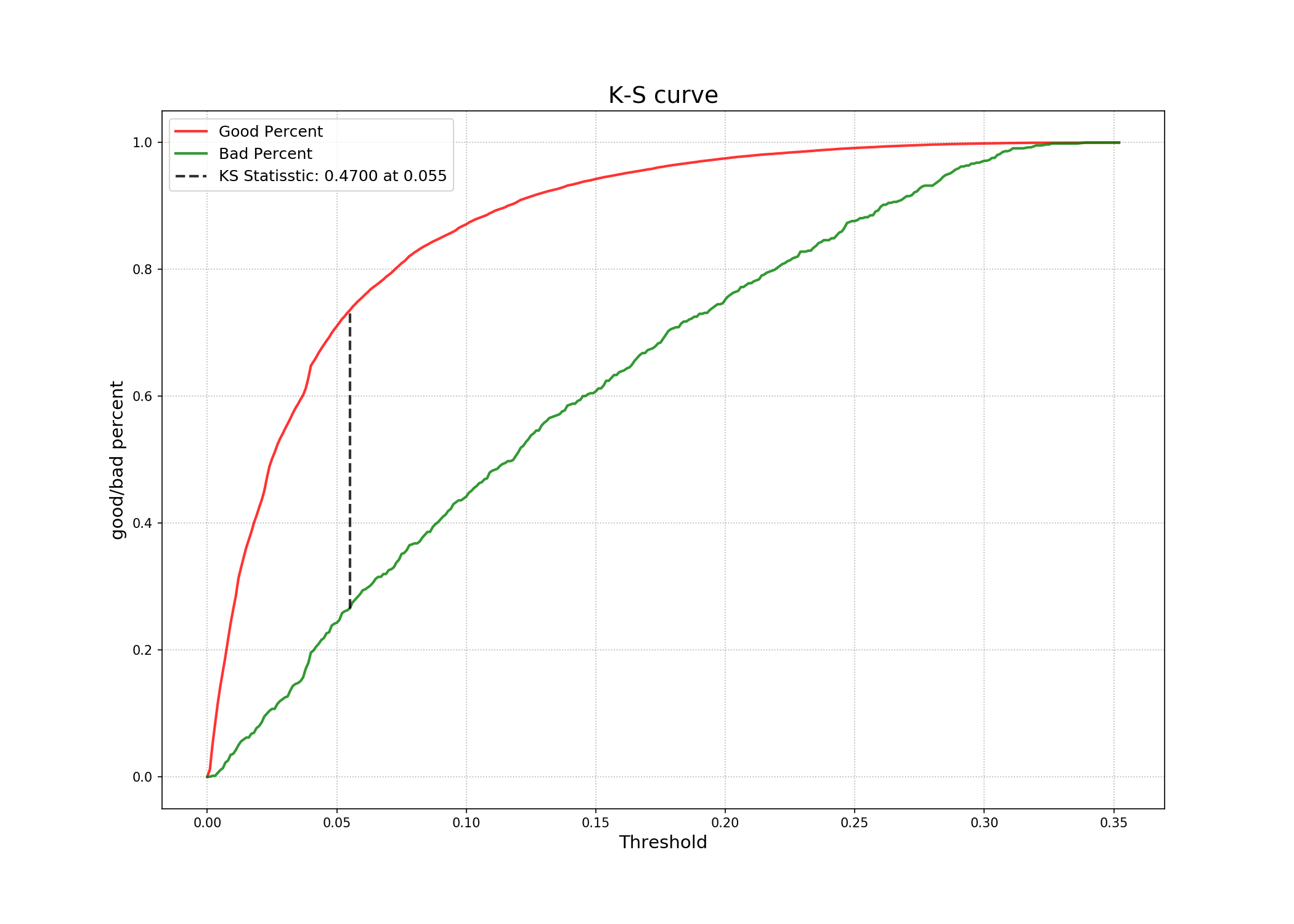
表2：模型基本指标表现

模型的KS与AUC在测试集TEST\_01\_02和测试集TEST\_03\_05上的表现与训练集上有较大的差异。

# 模型KS与AUC图像

1，训练集

模型在训练集（6-12月份）上的KS与AUC 图像如下：

 图4：训练集6-12月份KS曲线 未加正态分布

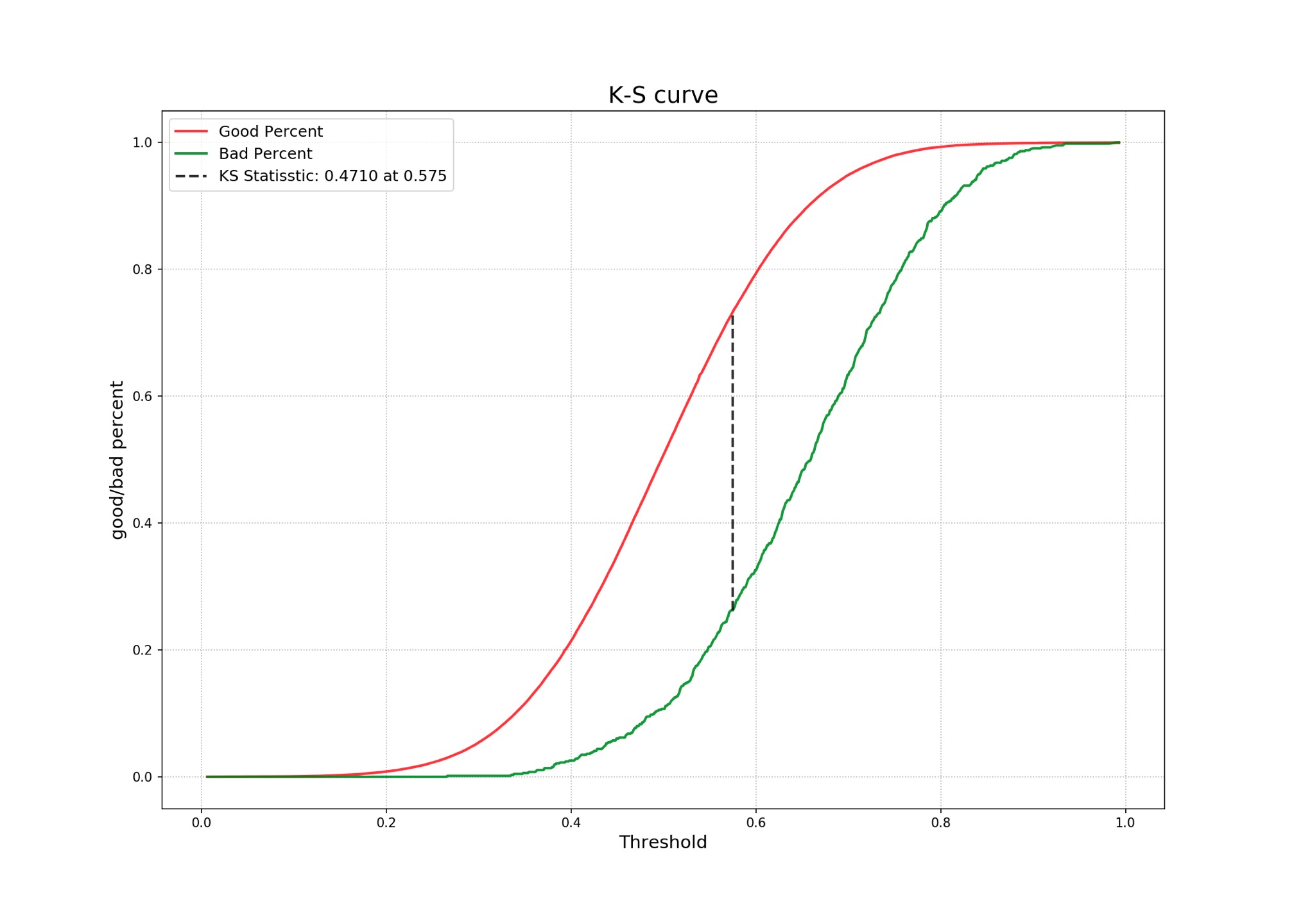


图5：训练集6-12月份KS曲线 加正态分布

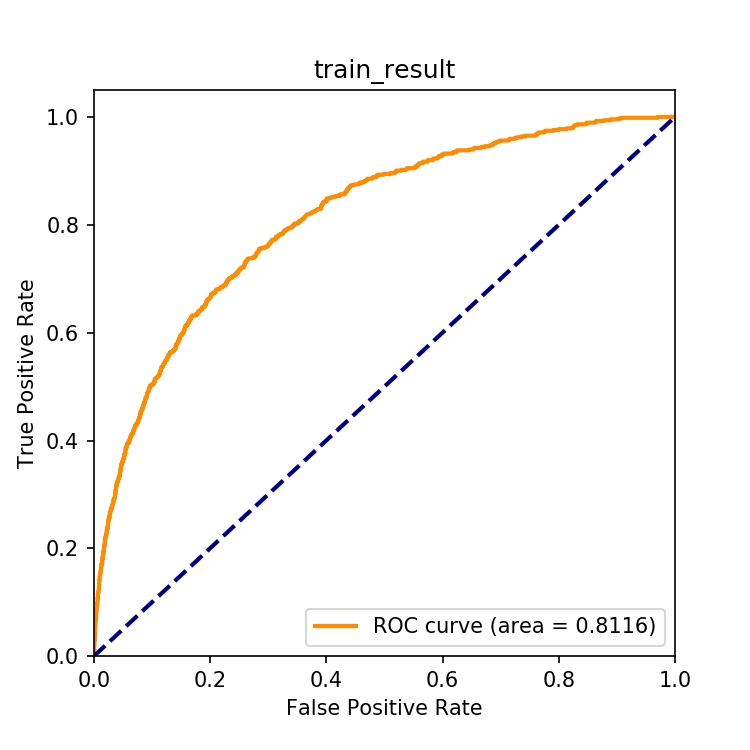
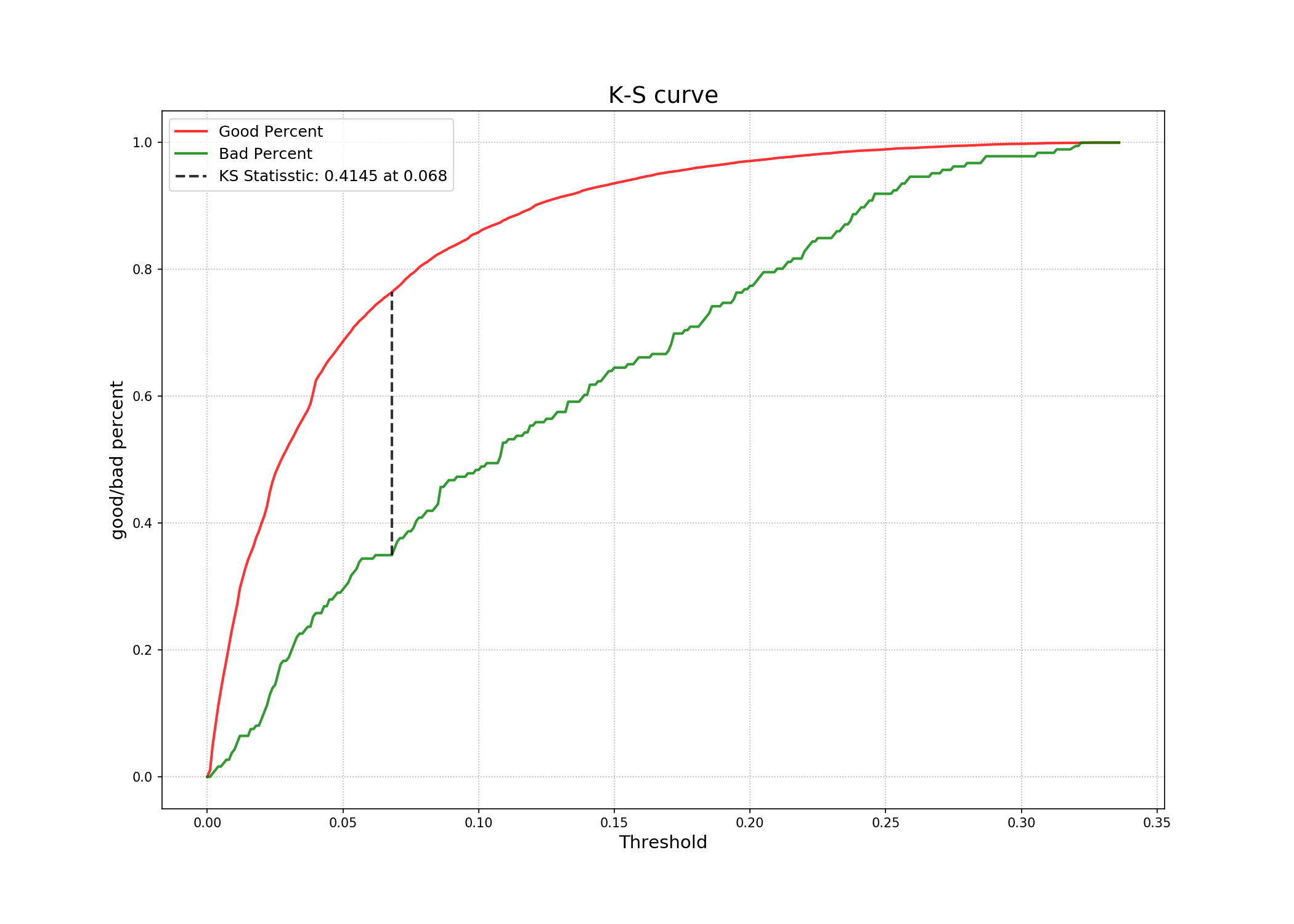


图6：训练集6-12月份AUC曲线

2，测试集1

模型在测试集 1 （2017年1月至2月的数据）上的KS与AUC 图像如下：

 图7：测试集1-2月份KS曲线 未加正态分布

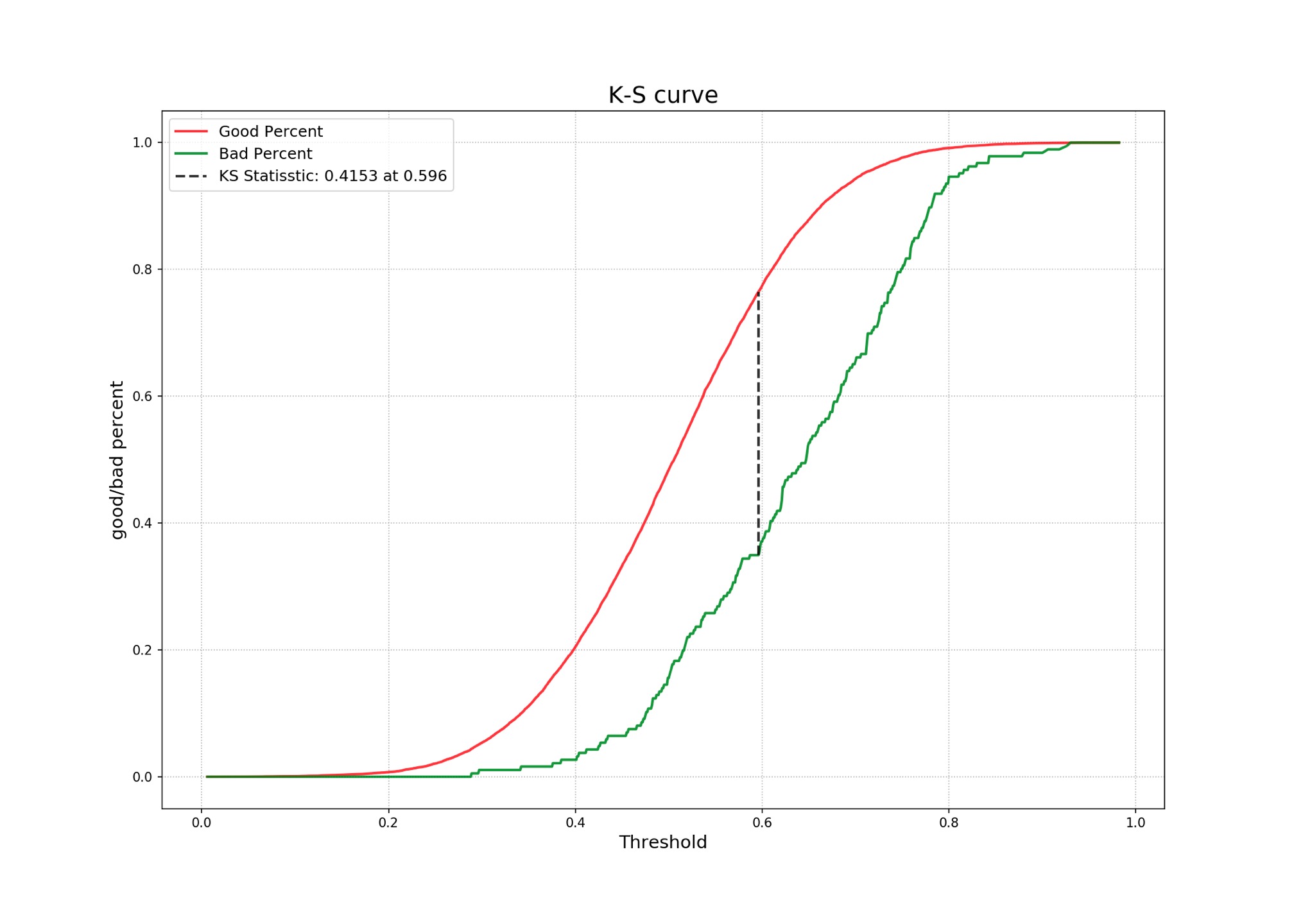


图8：测试集1-2月份KS曲线 加正态分布

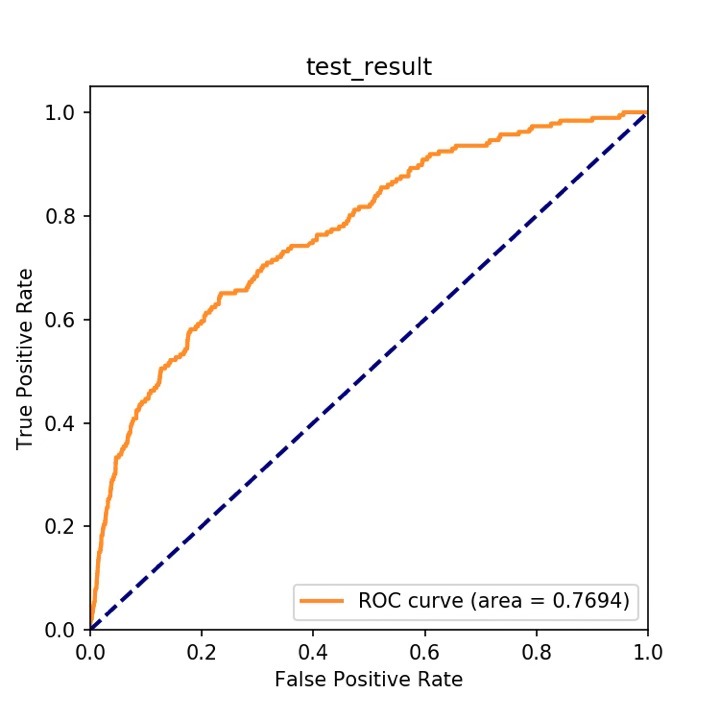


图9：测试集1-2月份AUC曲线

# 模型输出分布对比及PSI

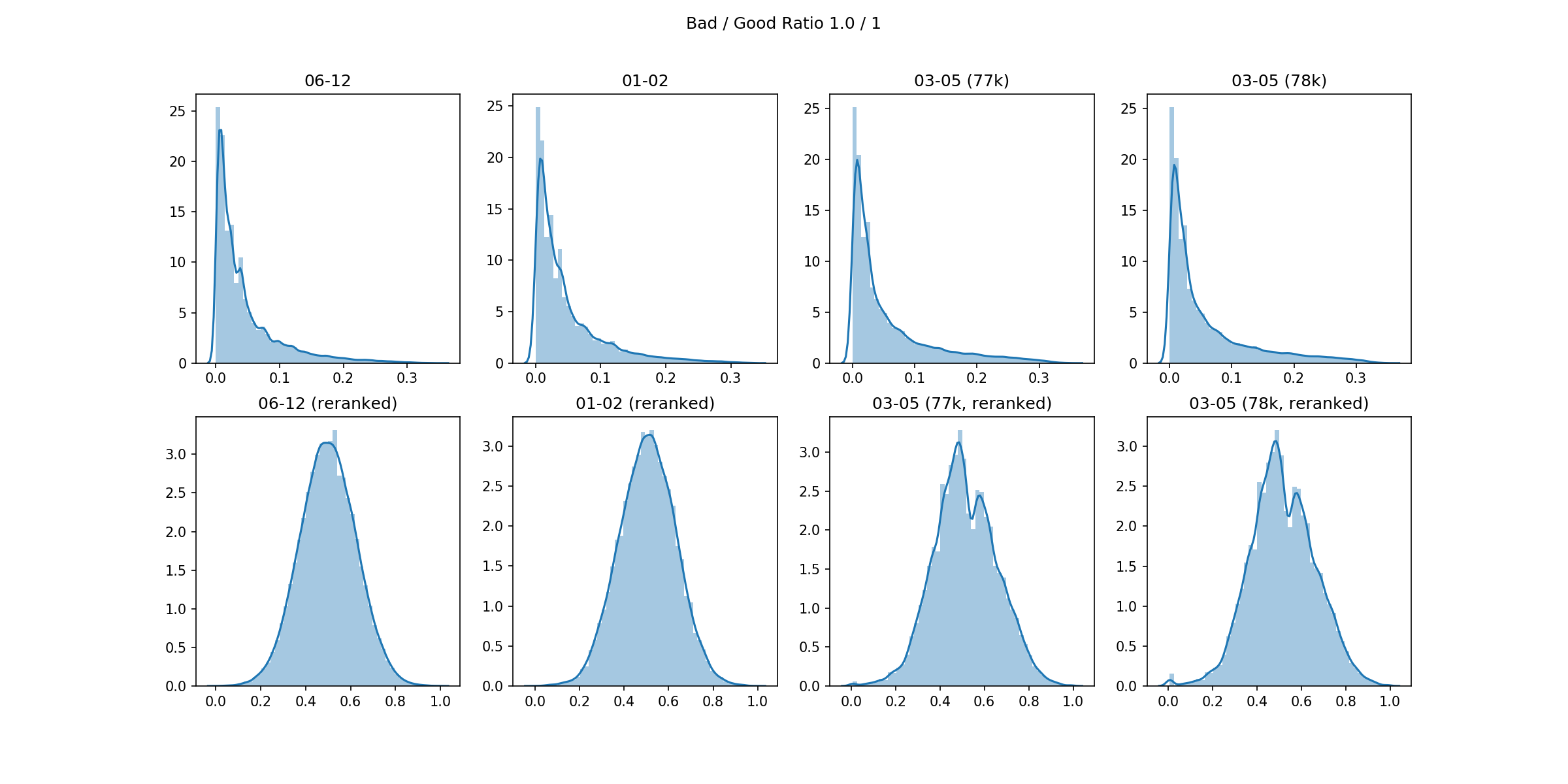
模型在训练集与测试集上的输出分布如下：

图10：模型概率分布图

|  |  |  |  |
| --- | --- | --- | --- |
| BIN | TEST\_01\_02 | TEST\_03\_05 | TEST\_03\_05(V2) |
| [0.894, inf) | 0.00001 | 0.00310 | 0.00537 |
| [0.796, 0.894) | 0.00014 | 0.01383 | 0.01921 |
| [0.697, 0.796) | 0.00047 | 0.01789 | 0.02127 |
| [0.598, 0.697) | 0.00101 | 0.00081 | 0.00080 |
| [0.500, 0.598) | 0.00004 | 0.00723 | 0.00815 |
| [0.401, 0.500) | 0.00064 | 0.00028 | 0.00066 |
| [0.302, 0.401) | 0.00044 | 0.00126 | 0.00147 |
| [0.204, 0.302) | 0.00000 | 0.00001 | 0.00004 |
| [0.105, 0.204) | 0.00024 | 0.00105 | 0.00092 |
| [0.000, 0.105) | 0.00007 | 0.00350 | 0.00828 |
| 总计 | 0.00307 | 0.04896 | 0.06614 |

表3：模型PSI

# 分布宽表

训练集

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PER\_BIN | 坏客户 | 好客户 | 放款 | 放款覆盖度 | DPD |
| [ 0.98 , 1.0 ) | 0 | 2 | 2 | 0.00% | 0.00% |
| [ 0.96 , 0.98 ) | 0 | 5 | 5 | 0.01% | 0.00% |
| [ 0.94 , 0.96 ) | 0 | 8 | 8 | 0.02% | 0.00% |
| [ 0.92 , 0.94 ) | 0 | 15 | 15 | 0.04% | 0.00% |
| [ 0.9 , 0.92 ) | 0 | 18 | 18 | 0.06% | 0.00% |
| [ 0.88 , 0.9 ) | 0 | 47 | 47 | 0.12% | 0.00% |
| [ 0.86 , 0.88 ) | 0 | 67 | 67 | 0.20% | 0.00% |
| [ 0.84 , 0.86 ) | 0 | 91 | 91 | 0.32% | 0.00% |
| [ 0.82 , 0.84 ) | 0 | 160 | 160 | 0.52% | 0.00% |
| [ 0.8 , 0.82 ) | 0 | 232 | 232 | 0.82% | 0.00% |
| [ 0.78 , 0.8 ) | 0 | 338 | 338 | 1.24% | 0.00% |
| [ 0.76 , 0.78 ) | 0 | 451 | 451 | 1.81% | 0.00% |
| [ 0.74 , 0.76 ) | 0 | 669 | 669 | 2.66% | 0.00% |
| [ 0.72 , 0.74 ) | 1 | 896 | 897 | 3.79% | 0.03% |
| [ 0.7 , 0.72 ) | 0 | 1223 | 1223 | 5.34% | 0.02% |
| [ 0.68 , 0.7 ) | 0 | 1603 | 1603 | 7.37% | 0.02% |
| [ 0.66 , 0.68 ) | 2 | 2041 | 2043 | 9.95% | 0.04% |
| [ 0.64 , 0.66 ) | 2 | 2491 | 2493 | 13.11% | 0.05% |
| [ 0.62 , 0.64 ) | 5 | 2986 | 2991 | 16.89% | 0.07% |
| [ 0.6 , 0.62 ) | 7 | 3380 | 3387 | 21.18% | 0.10% |
| [ 0.58 , 0.6 ) | 7 | 3947 | 3954 | 26.18% | 0.12% |
| [ 0.56 , 0.58 ) | 11 | 4378 | 4389 | 31.73% | 0.14% |
| [ 0.54 , 0.56 ) | 8 | 4719 | 4727 | 37.71% | 0.14% |
| [ 0.52 , 0.54 ) | 17 | 4954 | 4971 | 44.00% | 0.17% |
| [ 0.5 , 0.52 ) | 11 | 4957 | 4968 | 50.28% | 0.18% |
| [ 0.48 , 0.5 ) | 24 | 5008 | 5032 | 56.65% | 0.21% |
| [ 0.46 , 0.48 ) | 25 | 4970 | 4995 | 62.96% | 0.24% |
| [ 0.44 , 0.46 ) | 31 | 4489 | 4520 | 68.68% | 0.28% |
| [ 0.42 , 0.44 ) | 34 | 4204 | 4238 | 74.04% | 0.32% |
| [ 0.4 , 0.42 ) | 31 | 3795 | 3826 | 78.88% | 0.35% |
| [ 0.38 , 0.4 ) | 36 | 3425 | 3461 | 83.26% | 0.38% |
| [ 0.36 , 0.38 ) | 44 | 2923 | 2967 | 87.01% | 0.43% |
| [ 0.34 , 0.36 ) | 37 | 2365 | 2402 | 90.05% | 0.47% |
| [ 0.32 , 0.34 ) | 47 | 1954 | 2001 | 92.58% | 0.52% |
| [ 0.3 , 0.32 ) | 40 | 1543 | 1583 | 94.59% | 0.56% |
| [ 0.28 , 0.3 ) | 45 | 1147 | 1192 | 96.09% | 0.61% |
| [ 0.26 , 0.28 ) | 31 | 910 | 941 | 97.28% | 0.64% |
| [ 0.24 , 0.26 ) | 39 | 671 | 710 | 98.18% | 0.69% |
| [ 0.22 , 0.24 ) | 28 | 470 | 498 | 98.81% | 0.72% |
| [ 0.2 , 0.22 ) | 28 | 299 | 327 | 99.23% | 0.75% |
| [ 0.18 , 0.2 ) | 21 | 214 | 235 | 99.52% | 0.78% |
| [ 0.16 , 0.18 ) | 17 | 118 | 135 | 99.69% | 0.80% |
| [ 0.14 , 0.16 ) | 13 | 82 | 95 | 99.81% | 0.81% |
| [ 0.12 , 0.14 ) | 8 | 49 | 57 | 99.89% | 0.82% |
| [ 0.1 , 0.12 ) | 7 | 25 | 32 | 99.93% | 0.83% |
| [ 0.08 , 0.1 ) | 1 | 27 | 28 | 99.96% | 0.83% |
| [ 0.06 , 0.08 ) | 4 | 11 | 15 | 99.98% | 0.84% |
| [ 0.04 , 0.06 ) | 0 | 7 | 7 | 99.99% | 0.84% |
| [ 0.02 , 0.04 ) | 0 | 4 | 4 | 99.99% | 0.84% |
| [ 0.0 , 0.02 ) | 1 | 3 | 4 | 100.00% | 0.84% |
| 总计 | 663 | 78391 | 79054 |  |  |

Table 4 训练集6-12月份上输出分布宽表

测试集1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PER\_BIN | 坏客户 | 好客户 | 放款 | 放款覆盖度 | DPD |
| [ 0.98 , 1.0 ) | 0 | 1 | 1 | 0.00% | 0.00% |
| [ 0.96 , 0.98 ) | 0 | 1 | 1 | 0.01% | 0.00% |
| [ 0.94 , 0.96 ) | 0 | 5 | 5 | 0.03% | 0.00% |
| [ 0.92 , 0.94 ) | 0 | 13 | 13 | 0.07% | 0.00% |
| [ 0.9 , 0.92 ) | 0 | 6 | 6 | 0.09% | 0.00% |
| [ 0.88 , 0.9 ) | 0 | 18 | 18 | 0.16% | 0.00% |
| [ 0.86 , 0.88 ) | 0 | 26 | 26 | 0.25% | 0.00% |
| [ 0.84 , 0.86 ) | 0 | 30 | 30 | 0.36% | 0.00% |
| [ 0.82 , 0.84 ) | 0 | 41 | 41 | 0.51% | 0.00% |
| [ 0.8 , 0.82 ) | 0 | 58 | 58 | 0.72% | 0.00% |
| [ 0.78 , 0.8 ) | 0 | 121 | 121 | 1.15% | 0.00% |
| [ 0.76 , 0.78 ) | 0 | 131 | 131 | 1.63% | 0.00% |
| [ 0.74 , 0.76 ) | 0 | 252 | 252 | 2.54% | 0.00% |
| [ 0.72 , 0.74 ) | 0 | 324 | 324 | 3.70% | 0.00% |
| [ 0.7 , 0.72 ) | 2 | 447 | 449 | 5.32% | 0.14% |
| [ 0.68 , 0.7 ) | 0 | 535 | 535 | 7.25% | 0.10% |
| [ 0.66 , 0.68 ) | 0 | 651 | 651 | 9.60% | 0.08% |
| [ 0.64 , 0.66 ) | 1 | 846 | 847 | 12.66% | 0.09% |
| [ 0.62 , 0.64 ) | 1 | 1017 | 1018 | 16.33% | 0.09% |
| [ 0.6 , 0.62 ) | 1 | 1087 | 1088 | 20.25% | 0.09% |
| [ 0.58 , 0.6 ) | 3 | 1293 | 1296 | 24.92% | 0.12% |
| [ 0.56 , 0.58 ) | 4 | 1455 | 1459 | 30.19% | 0.14% |
| [ 0.54 , 0.56 ) | 2 | 1488 | 1490 | 35.56% | 0.14% |
| [ 0.52 , 0.54 ) | 6 | 1673 | 1679 | 41.61% | 0.17% |
| [ 0.5 , 0.52 ) | 9 | 1776 | 1785 | 48.05% | 0.22% |
| [ 0.48 , 0.5 ) | 11 | 1681 | 1692 | 54.15% | 0.27% |
| [ 0.46 , 0.48 ) | 8 | 1825 | 1833 | 60.77% | 0.28% |
| [ 0.44 , 0.46 ) | 5 | 1583 | 1588 | 66.49% | 0.29% |
| [ 0.42 , 0.44 ) | 11 | 1505 | 1516 | 71.96% | 0.32% |
| [ 0.4 , 0.42 ) | 5 | 1405 | 1410 | 77.04% | 0.32% |
| [ 0.38 , 0.4 ) | 10 | 1299 | 1309 | 81.77% | 0.35% |
| [ 0.36 , 0.38 ) | 12 | 1134 | 1146 | 85.90% | 0.38% |
| [ 0.34 , 0.36 ) | 10 | 913 | 923 | 89.23% | 0.41% |
| [ 0.32 , 0.34 ) | 9 | 732 | 741 | 91.90% | 0.43% |
| [ 0.3 , 0.32 ) | 11 | 581 | 592 | 94.03% | 0.46% |
| [ 0.28 , 0.3 ) | 11 | 436 | 447 | 95.65% | 0.50% |
| [ 0.26 , 0.28 ) | 12 | 346 | 358 | 96.94% | 0.54% |
| [ 0.24 , 0.26 ) | 12 | 268 | 280 | 97.95% | 0.57% |
| [ 0.22 , 0.24 ) | 11 | 192 | 203 | 98.68% | 0.61% |
| [ 0.2 , 0.22 ) | 8 | 109 | 117 | 99.10% | 0.64% |
| [ 0.18 , 0.2 ) | 3 | 86 | 89 | 99.42% | 0.65% |
| [ 0.16 , 0.18 ) | 2 | 45 | 47 | 99.59% | 0.65% |
| [ 0.14 , 0.16 ) | 2 | 48 | 50 | 99.77% | 0.66% |
| [ 0.12 , 0.14 ) | 0 | 21 | 21 | 99.85% | 0.66% |
| [ 0.1 , 0.12 ) | 1 | 16 | 17 | 99.91% | 0.66% |
| [ 0.08 , 0.1 ) | 1 | 10 | 11 | 99.95% | 0.66% |
| [ 0.06 , 0.08 ) | 2 | 8 | 10 | 99.99% | 0.67% |
| [ 0.04 , 0.06 ) | 0 | 3 | 3 | 100.00% | 0.67% |
| [ 0.02 , 0.04 ) | 0 | 0 | 0 | 100.00% | 0.67% |
| [ 0.0 , 0.02 ) | 0 | 1 | 1 | 100.00% | 0.67% |
| 总计 | 186 | 27542 | 27728 |  |  |

Table 5 测试集1-2月份上输出分布宽表