同济大学 TONGJI UNIVERSITY

面向磁浮轨道异常检测的大数据分析框架研究

小组汇报

2024-11-19　　By 刘震

基本思路：

1.扩充数据集，尤其是明显异常点，数据样本数量严重不平衡

1.1欠采样、过采样特定类的方法

1.2加权损失函数

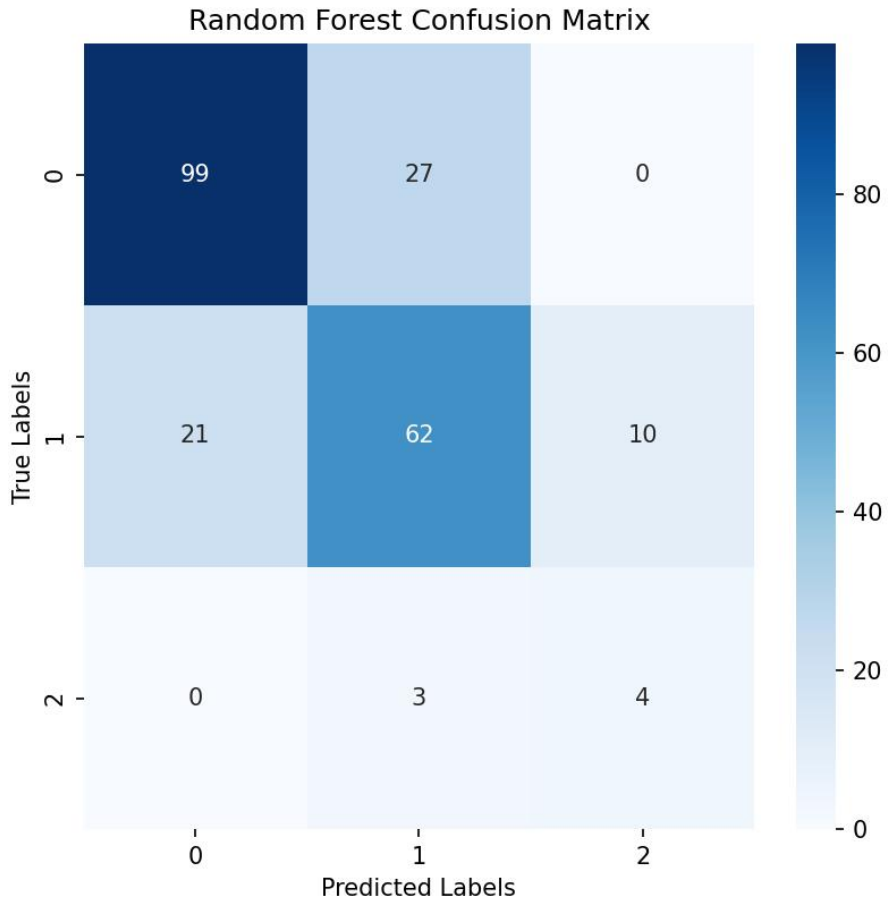
1.3调整阈值设定

2.融合算法模型

基于SMOTE对数据集过采样：

特点：对标签2数据的准确率很低

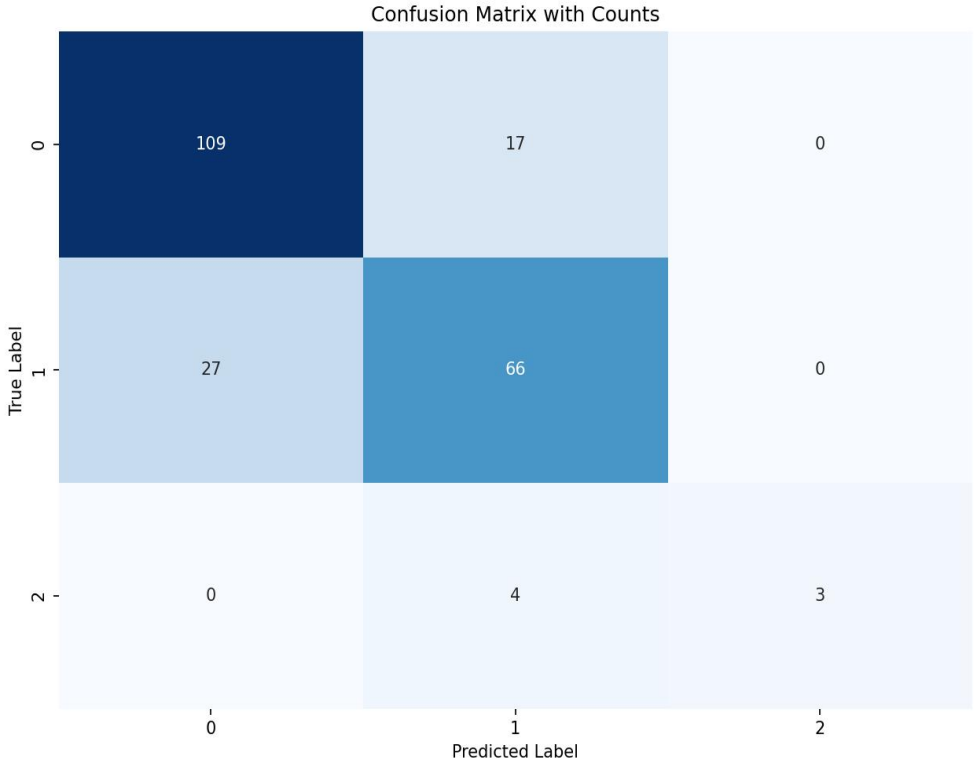
标签0、1之前难以区分

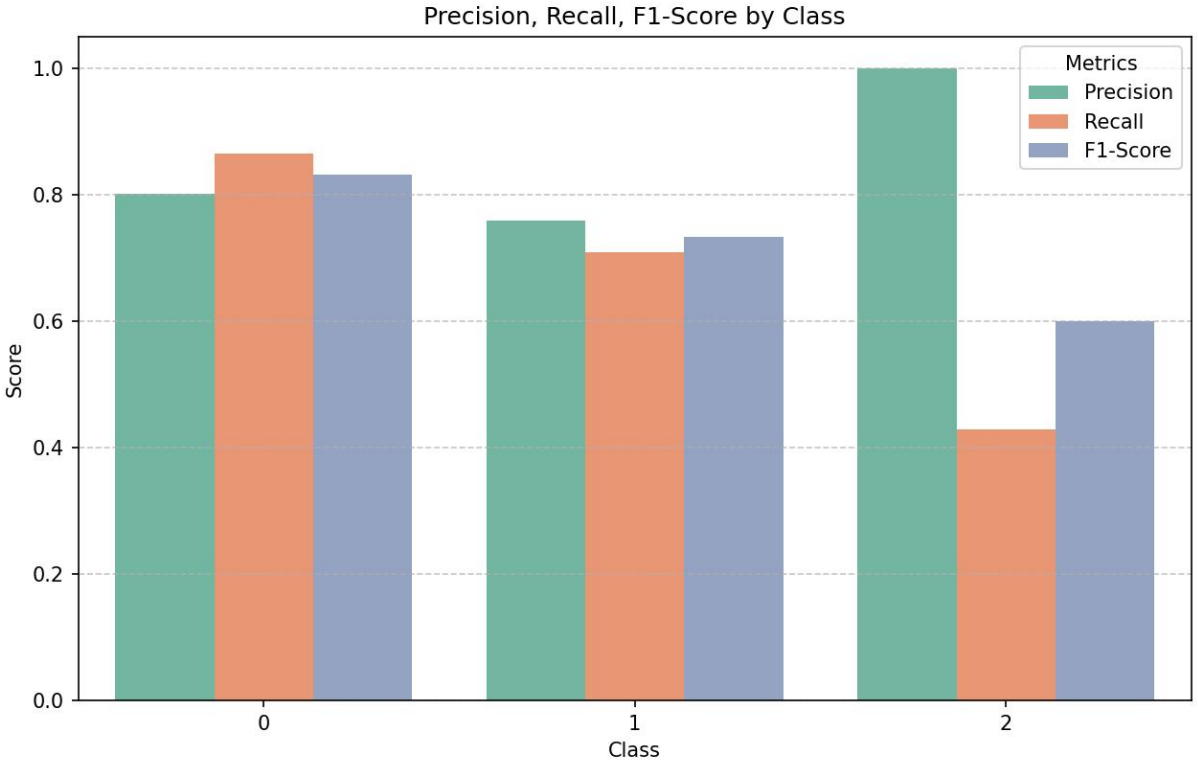


Weighted F1 Score: 0.7364

基于RF，对损失函数进行加权处理，针对样本数据不平衡的问题

$$\text{Loss} = \frac{1}{\text{TotalSamples}} \sum_{i=1}^{\text{TotalSamples}} \text{ClassWeight}[y_i] \cdot \text{SampleLoss}(h(x_i), y_i)$$

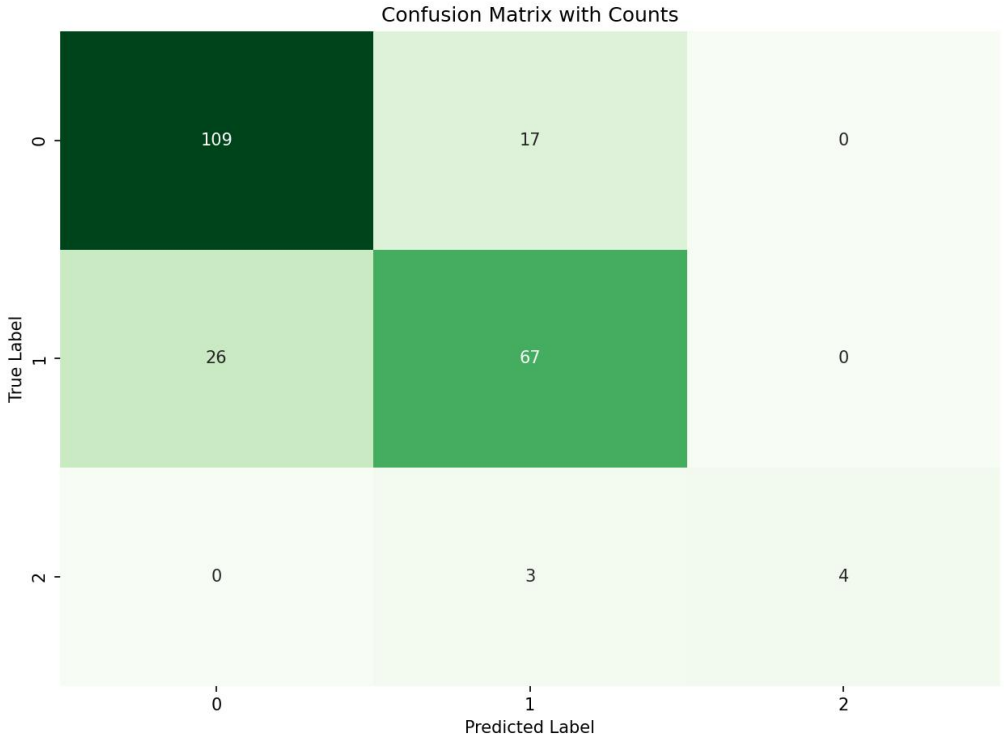


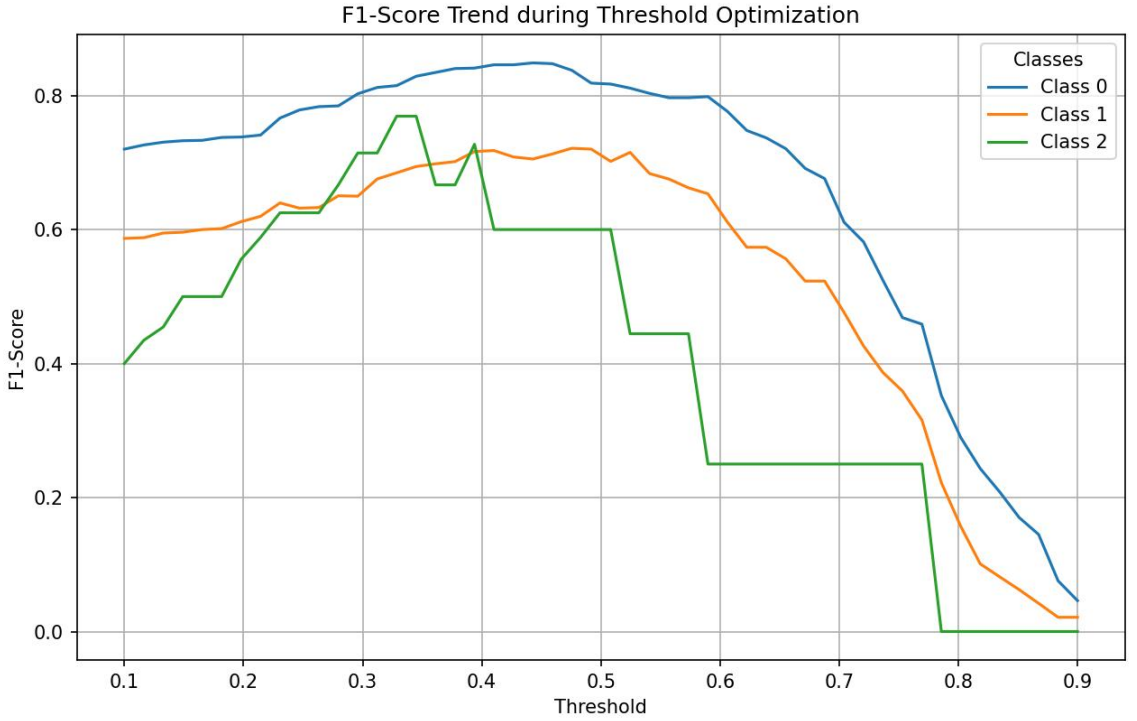


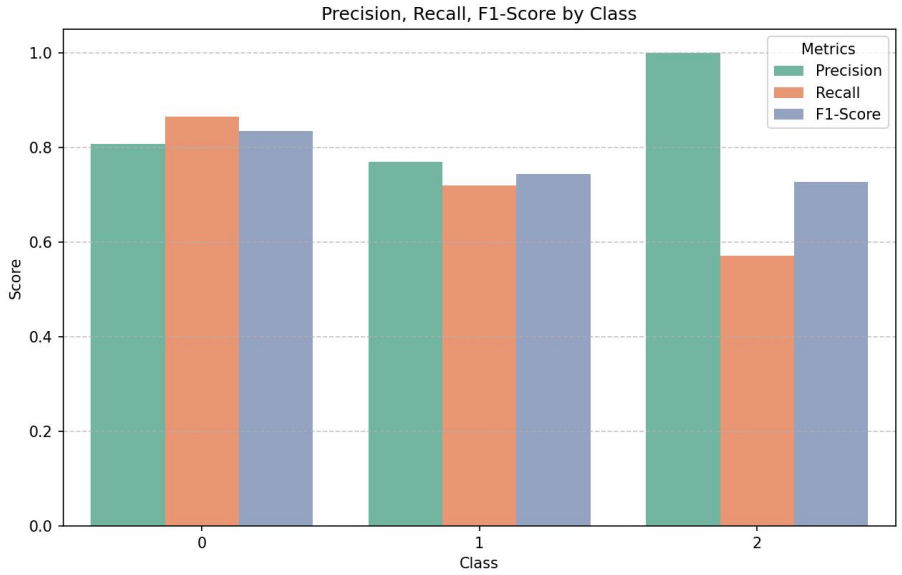
Weighted F1-Score: 0.7836

进行阈值优化，而不是简单的比较谁的预测概率值更大，即：

从0到1之间分别找出每个类的阈值，能让整体F1-score最大

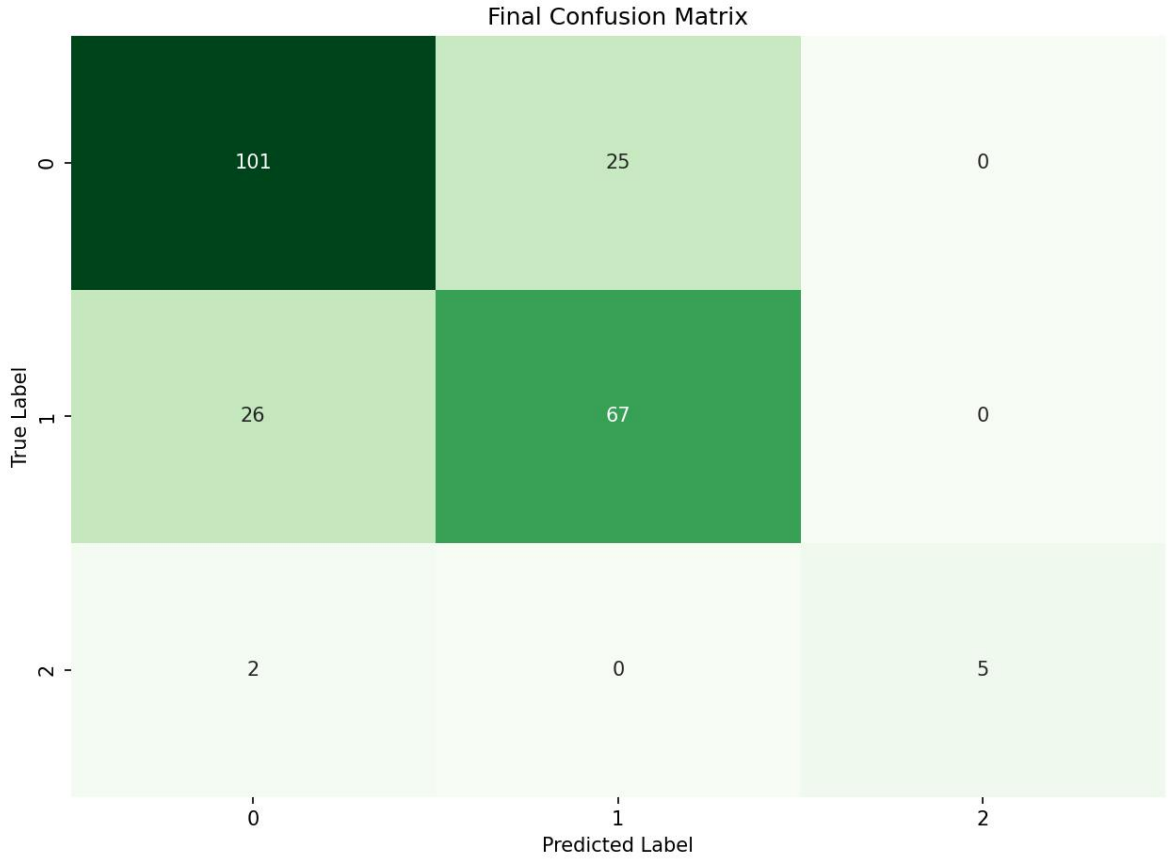






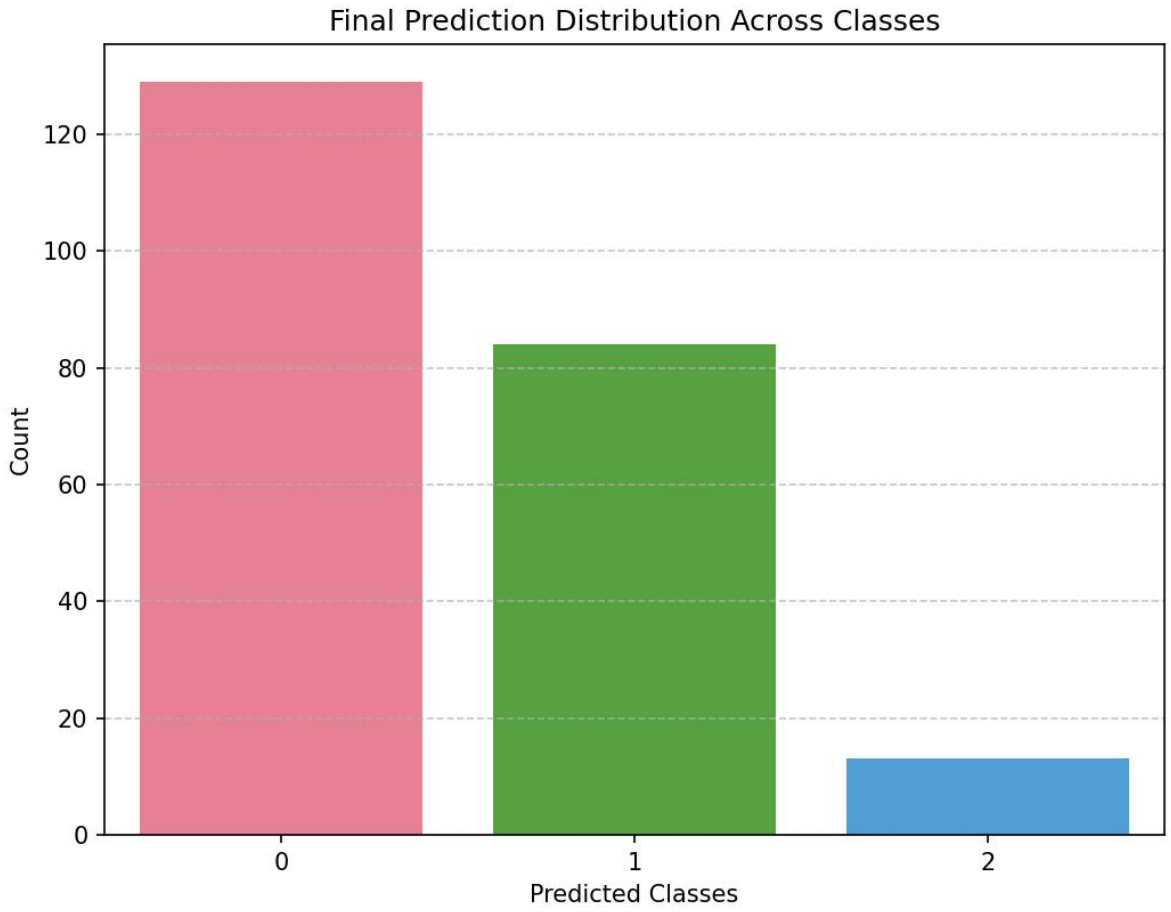
Weighted F1-Score: 0.7945 （+1.4%）

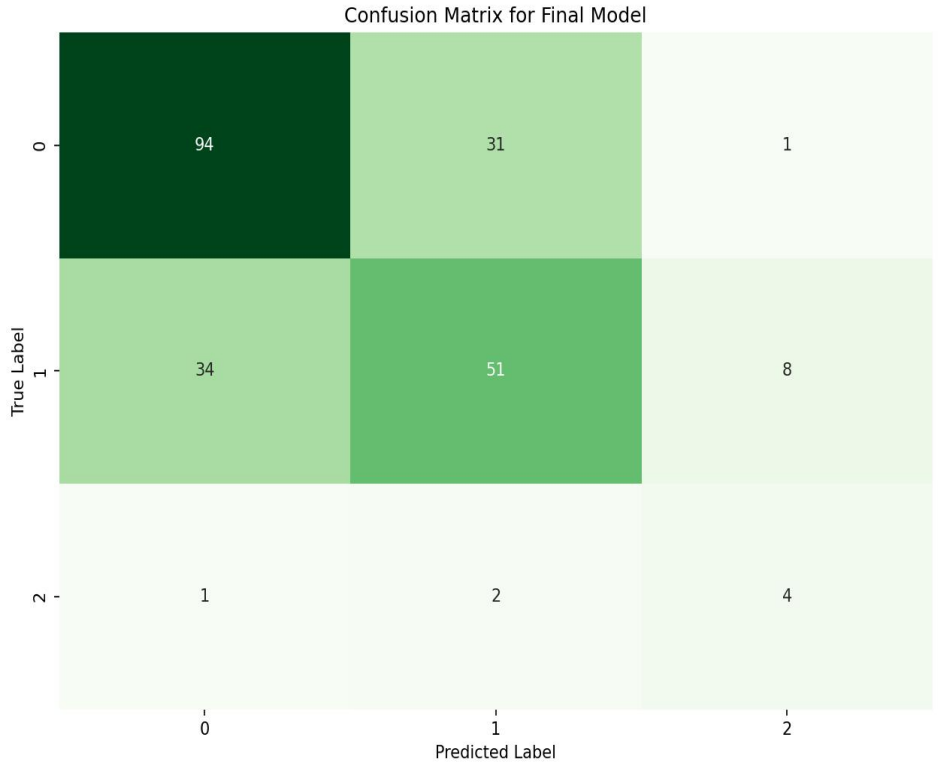
为了区分0、1标签的数据，可以分步检测，即先检测标签为2的，将0，1当作正常点，再将2剔除后检测0，1。然而，这种方式的效果并没有之前的好，无论是MLP还是RF



Final Weighted F1-Score: 0.7655（-2.3%）

对算法模型的融合，即先用MLP和RF初始检测，再SMOTE扩充数据，进行过采样，用KNN检测，二者的结果进行投票加权





Final Weighted F1-Score: 0.7126

(显然结果相比于之前不是很好)

问题：

标签0和标签1的数据还是很难区分开来，后序可以想办法增加区分度

KNN算法不能很好的与其他算法融合，发挥其优势，可能存在代码或逻辑上的问题，也可能是方法使用不佳