不同模型恶意url识别系统

一、项目背景

随着互联网的飞速发展，恶意URL（如钓鱼链接、携带恶意软件的网站等）对网络安全构成了严重威胁。传统的黑名单或特征匹配方式在面对新型攻击时显得力不从心，因此，如何基于机器学习方法构建一个高效准确的恶意网址识别系统成为当前网络空间安全领域的重要研究方向。

本项目旨在构建一个结合文本特征与结构化特征的恶意网址识别系统，使用多种机器学习算法进行建模和对比，最终实现对未知恶意URL的自动识别。

二、数据集构建

我从以下数据源收集URL数据：

Openphish（恶意网址）

URLhaus（恶意网址）

Tranco（可信网站）

Kaggle上的公开数据集（补充可信/恶意网址）

使用数据：训练数据约12万，正常url和恶意url之比接近一比一

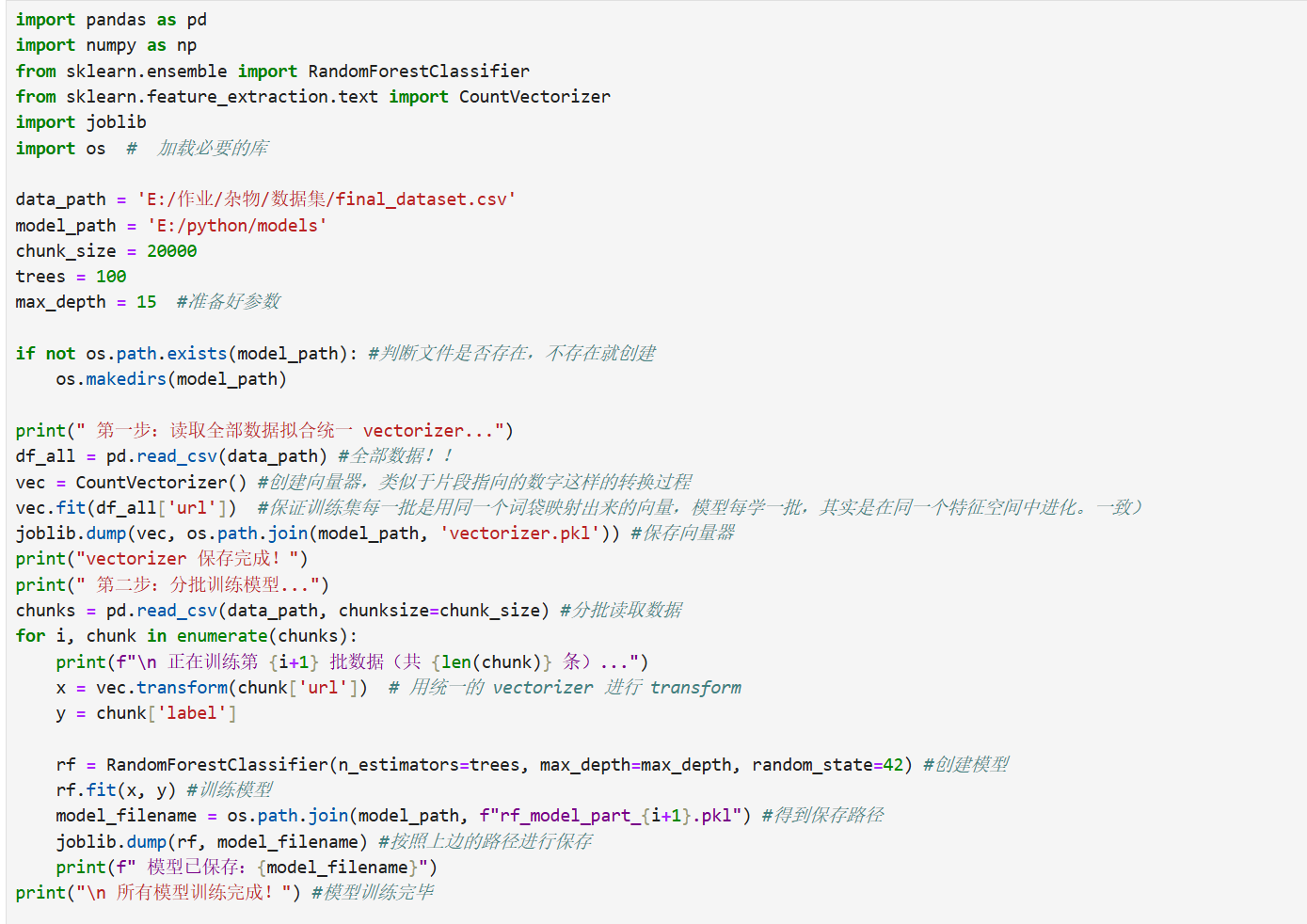
测试数据（并非一个数据集划分）单独成文件，数据约5万，正常url和恶意url之比也接近一比一（训练测试数据均已打乱）

（数据之比虽然不是常规的8:2，但7:3 或7.5:2.5比例也非常常见，不会响模型泛化能力，后边出于种种原因，我扩大了样本量训练24万，测试10万，数据均被打乱）

尝试使用随机森林

初始模型是按照12万数据进行训练，测试划分。（最开始使用的是随机森林，因为我在学这个）决策树设置100，但由于跑得太慢所以只使用了部分少量数据。结果导致准确率为1.00，宣告失败（没有模型能够预测到百分百对的）。

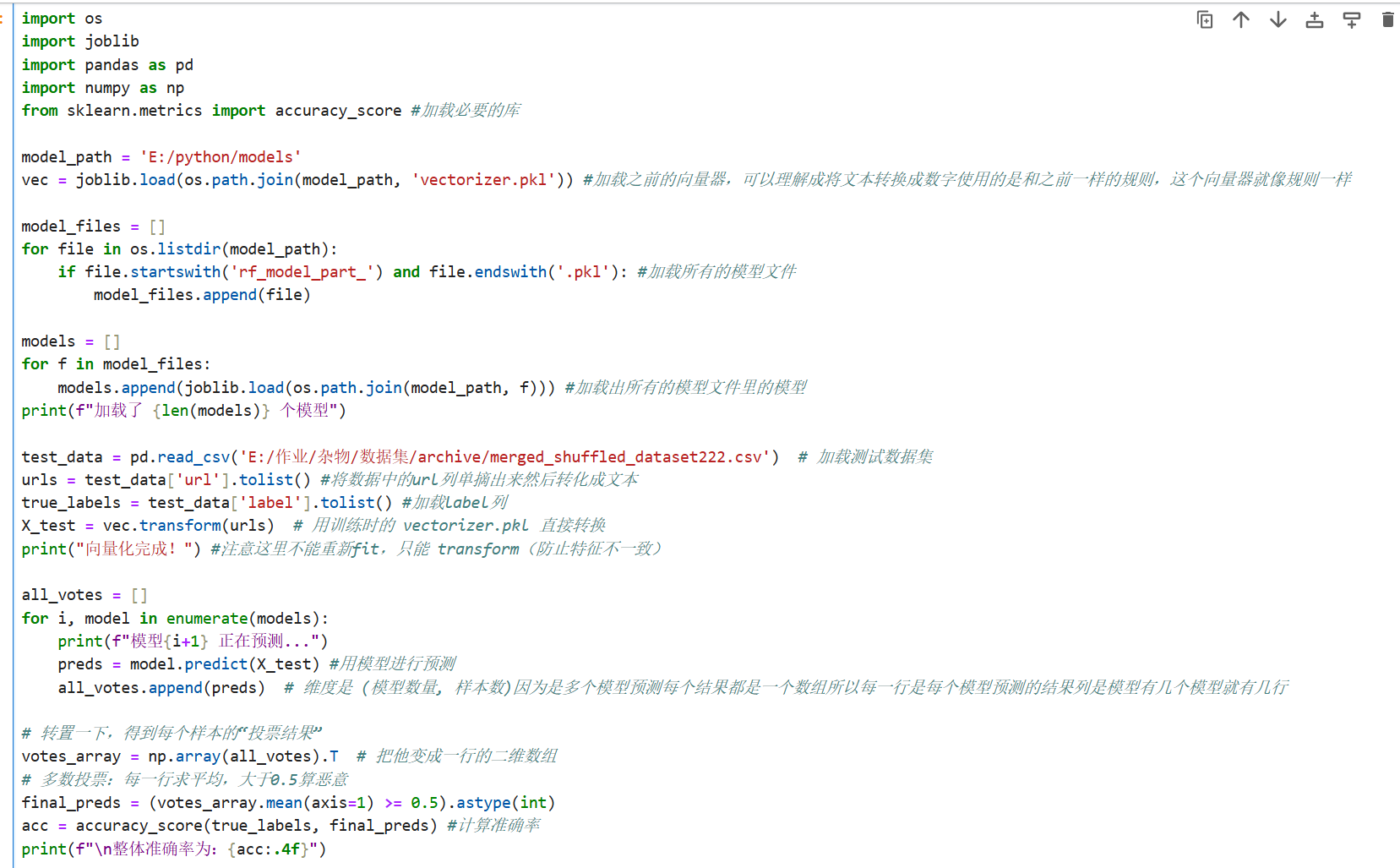
于是按照另一思路，通过**分批训练**多个模型最后再让他们一起预测跟据投票作出最终判断：（注释是我自己按自己理解写的）



运行结果是这样的：



然后就是使用测试数据集进行测试计算准确率：



最后结果是：



可以看到准确率只有0.38，后来发现测试集有问题我就更换测试集然后结果是

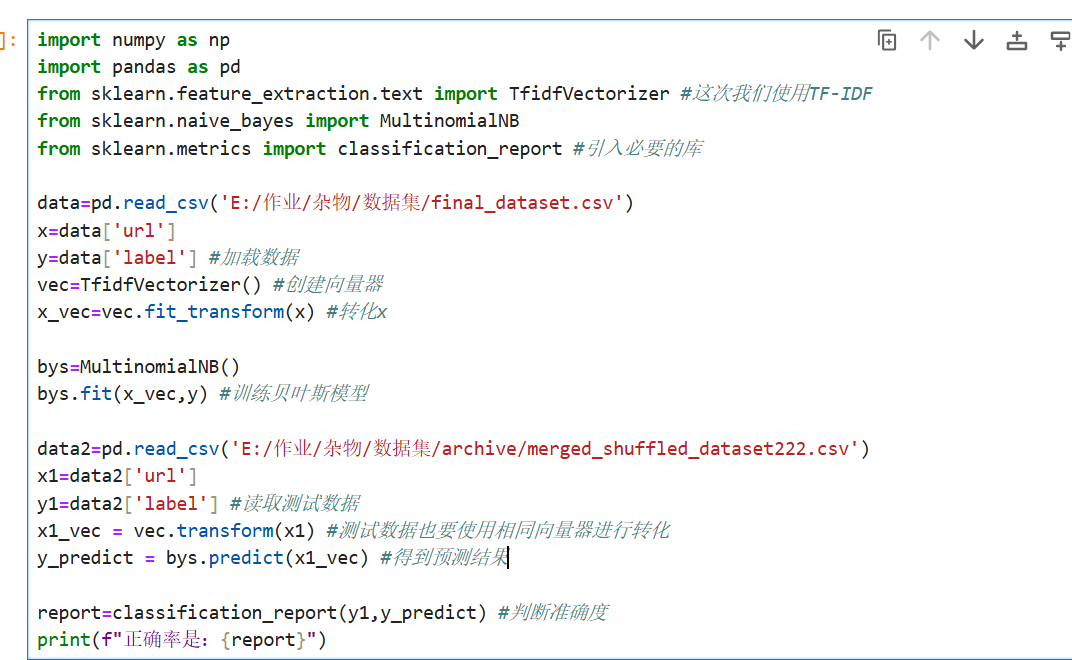


正确率在0.65左右，所以还是失败的，所以我们需要再次调整思路。

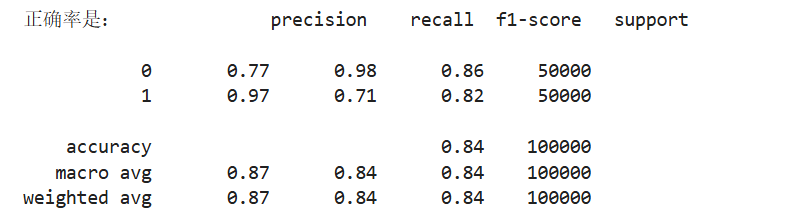
Tfidf + Naive Bayes尝试使用朴素贝叶斯

随机森林就是多个决策树类似二叉树形式慢慢追问判断，朴素贝叶斯就是直接根据词出现在恶意网站的概率来计算出这个网站是恶意网站的概率进而进行判断

代码如下：



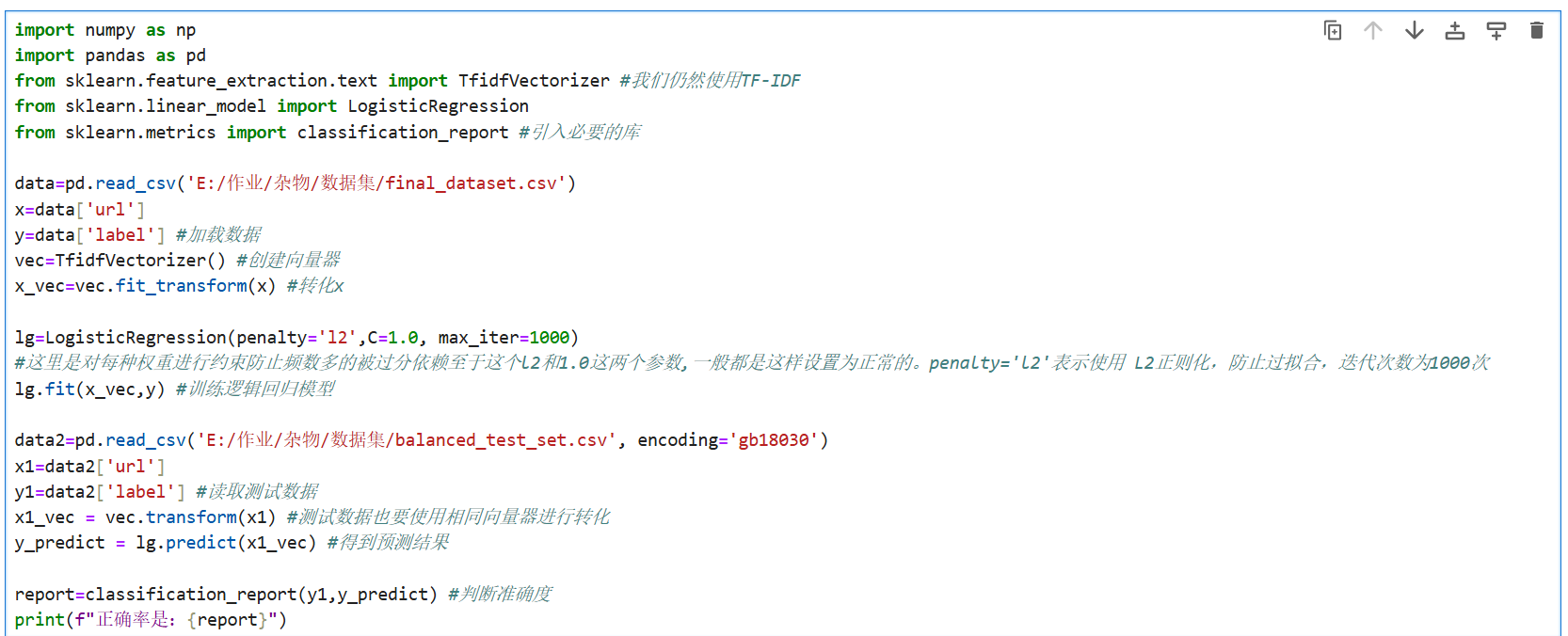
结果如下：



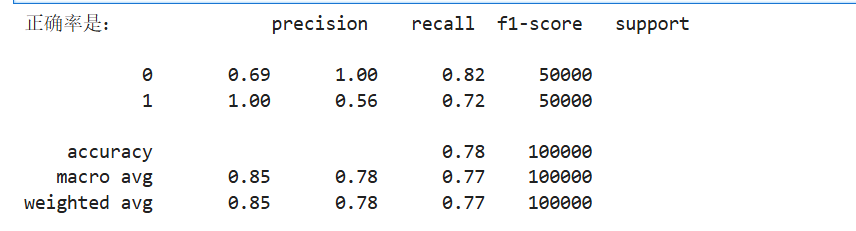
可以看出这次准确度在0.84，准确率有所上升但还是不够，我们需要进一步探讨。

逻辑回归模型

所以逻辑回归就是通过多次尝试找到每个特征的权重然后得到最终最为准确的参数设为公式进而进行预测。



和之前的模型相比，代码改动不大。结果是



正确率只有0.78左右还是不行，失败。。。。。。

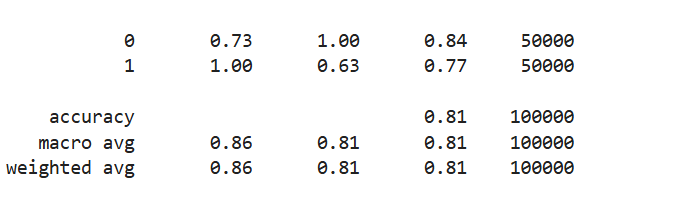
接下来从两个方向尝试，换模型或者扩大数据集

Kmeans是无监督的，对于我们这些标好的数据来说不太适合。Knn适合低维小量的数据，所以也不太适合。接下来尝试使用使用神经网络。

神经网络模型



最后结果是·



正确率在0.8，还是没有达到想要的样子。

尝试另一种，扩大样本量。我将训练数据扩大到了24万，正常和恶意都是12万。

1. 尝试使用随机森林分批次训练方法，准确率只有0.5，失败。
2. 尝试使用朴素贝叶斯模型，准确率只有0.8，未达到预期。
3. 尝试使用逻辑回归模型，准确率只有0.57，失败。
4. 尝试使用神经网络模型，准确率只有0.79，未达到预期。识别恶意网址的能力已经很强了，召回率高达 96%，**但误杀正常网址稍微有点多**（精确率 0.72）。

其实大部分都是预测的差不多，恶意网址召回率都很高，只是误杀率有点高。现在要做的就是优化模型，调参即可。

我尝试添加了少量的结构化特征：



1、尝试改进后的朴素贝叶斯模型，但是结果仍无明显变化。准确率还是0.8。

2、尝试改进后随机森林分批次训练方法，但是结果仍无明显变化。

3、尝试使用改进后神经网络模型，结果也无明显变化。能够看到模型宁可错杀一千绝不放过一个恶意的。唉……

尝试在改进后的神经网络模型基础上再加上权重的约束以改善这一情况：





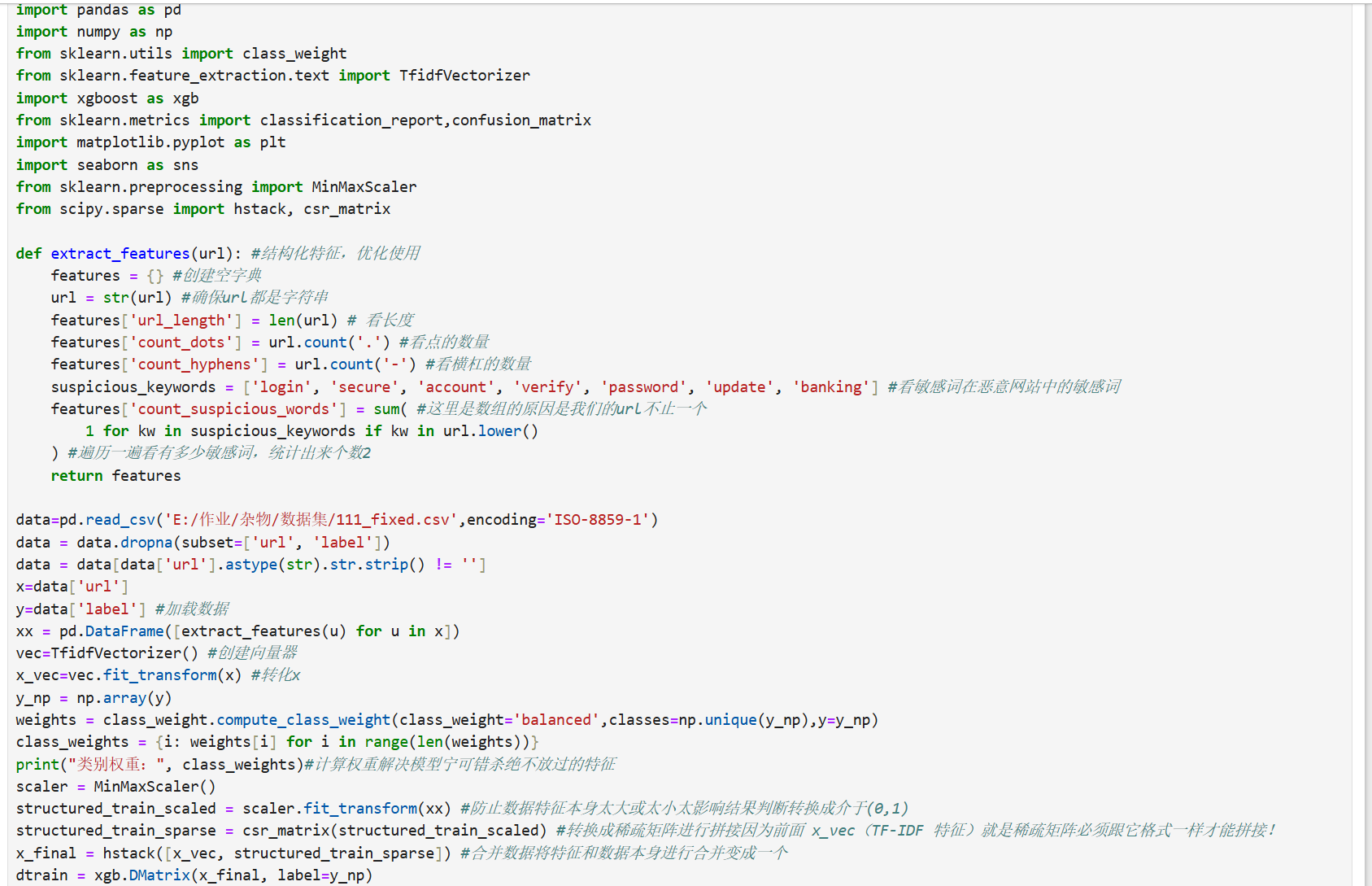
注释均是根据自己理解自行注释

但是结果仍未改善，效果仍然不理想。

XGBoost模型

XGBoost 是通过不断地**构建新树来纠正前面模型的错误**，最终形成一个强大的集成模型。

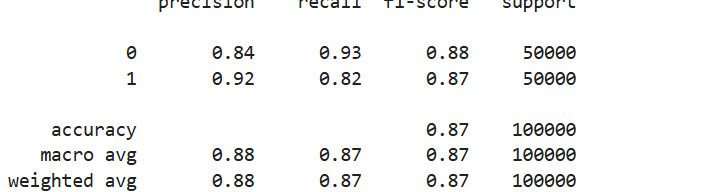
我们尝试使用XGBoost模型，代码改动不多：





结果如下：





可以看到结果已经在0.87，效果有所改善。

总结一下：

模型设计与实验

1、随机森林

准确率最高0.65，效果较差

2、朴素贝叶斯（MultinomialNB）

适用于文本类任务

准确率最高0.84，效果一般

3、逻辑回归（LogisticRegression）

对高维稀疏特征较友好

准确率最高0.78

4、神经网络（Keras构建）

两层全连接网络，64→32→1

加入class\_weight解决类别不均衡问题

准确率最高为0.81

5、 XGBoost模型（重点）

使用结构化+TFIDF特征

参数：max\_depth=8, eta=0.05, colsample\_bytree=0.8

准确率达到0.87，F1分数优异，召回率较高

模型融合探索（未完成）

尝试将XGBoost、神经网络、朴素贝叶斯和逻辑回归进行Stacking融合，计划以XGBoost为主模型，其他为辅助分类器。初步尝试效果不理想，暂未使用。

Transformer探索（实验阶段）

尝试基于字符级Token的Transformer模型进行网址分类，尽管Recall较高，但由于Precision偏低，整体准确率未超过0.45。未来可考虑更深层结构或结合结构化特征改进。

结果与总结

当前最优模型为XGBoost，准确率为0.87，表现稳定可靠。

未来优化方向：

引入更多结构化特征（如域名深度、特殊符号比例等）

模型融合策略完善，构建稳定的Stacking模型

探索Transformer与CNN结构的有效结合

自动化调参（GridSearch、Bayesian Optimization）

项目价值：

该系统具备实际落地能力，可用于对企业网关、邮件过滤器、浏览器插件等提供恶意URL识别支持。在当今网络安全及人工智能领域有着更加重要的作用。

以上均为个人制作（网络空间安全专业）

恳请批评指正