中國人民大學
RENMIN UNIVERSITY OF CHINA

财务报表欺诈检测

姓　　名：徐慧聪

学　　号：2019202363

学　　院：信息学院

班　　级：图灵二班

指导老师：许伟老师

财务报表欺诈检测

【摘要】

金融市场一直是世界中众多行业中发展最快，也是风险最大的行业之一，其中金融欺诈的方式和方法层出不穷，而财务报表的欺诈行为则是上市公司用以掩盖自身财务的重要手段。最近频繁发生的财务报表欺诈和舞弊案件震惊了社会公众,给投资者带来了重大损失,扰乱了证券市场的良性发展,对证券市场的健康发展形成重大危害。

本文根据印度 2011-2015 年之间的财务报表数据及欺诈标签，通过多种机器学习的方法建立模型，对财务报表欺诈行为进行识别，为财务报表欺诈的检测提供一种新的解决方案和思路。

【关键词】

财务欺诈；机器学习；欺诈检测

一、介绍

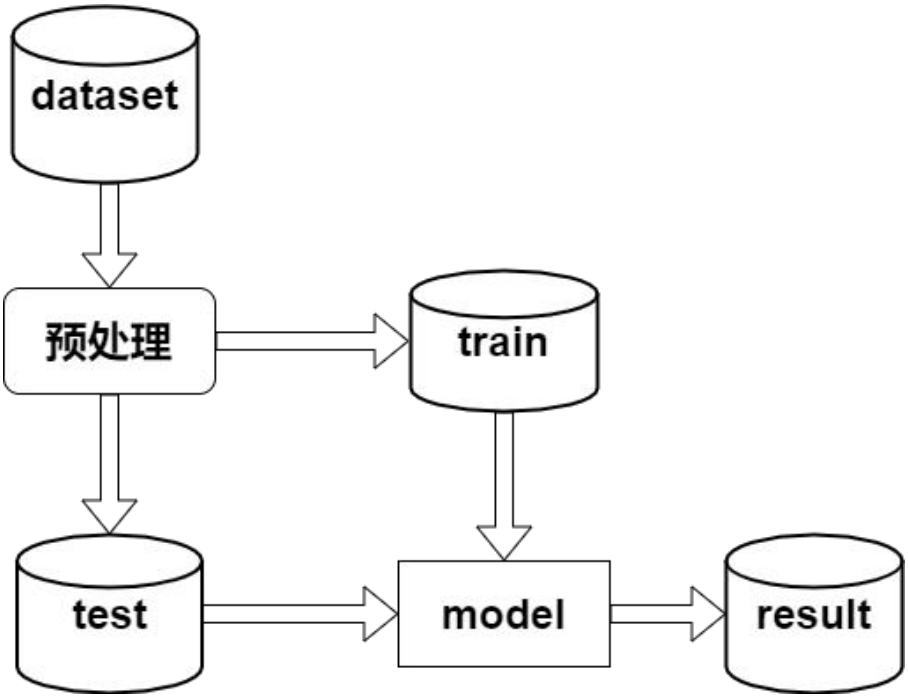
《中国注册会计师审计准则第 1141 号——财务报表审计中与舞弊相关的责任》将舞弊定义为被审计单位的管理层、治理层、员工或第三方使用欺骗手段获取不正当或非法利益的行为。因此，本文任务财务报表欺诈是指上市公司或会计主体使用不当方法使会计失真，故意谎报单位的经营成果和财务状况的违法违规行为。

本文的数据为印度 2011-2015 年间部分公司的财务报表，并根据印度公布的情况对欺诈情况进行了标注，之所以没有使用中国的数据是因为没有找到国内的相关数据集，数据集可以在 github 中获得（https://github.com/AdroMine/Predicting-Fraud-in-Financial-Statements）。

在本文中，我们根据数据集进行了多次实验，对不同模型之间进行对比，同时分析每一个特征对实验结果的影响程度。

二、思路

（1）架构



在本文，我们的架构主要分为三个部分，分别为数据集的获取和预处理，模型的训练和测试集的预测，最后结果的分析和再训练，之后我们会分别讲述不同部分我们的工作。

（2）数据集的获取和预处理

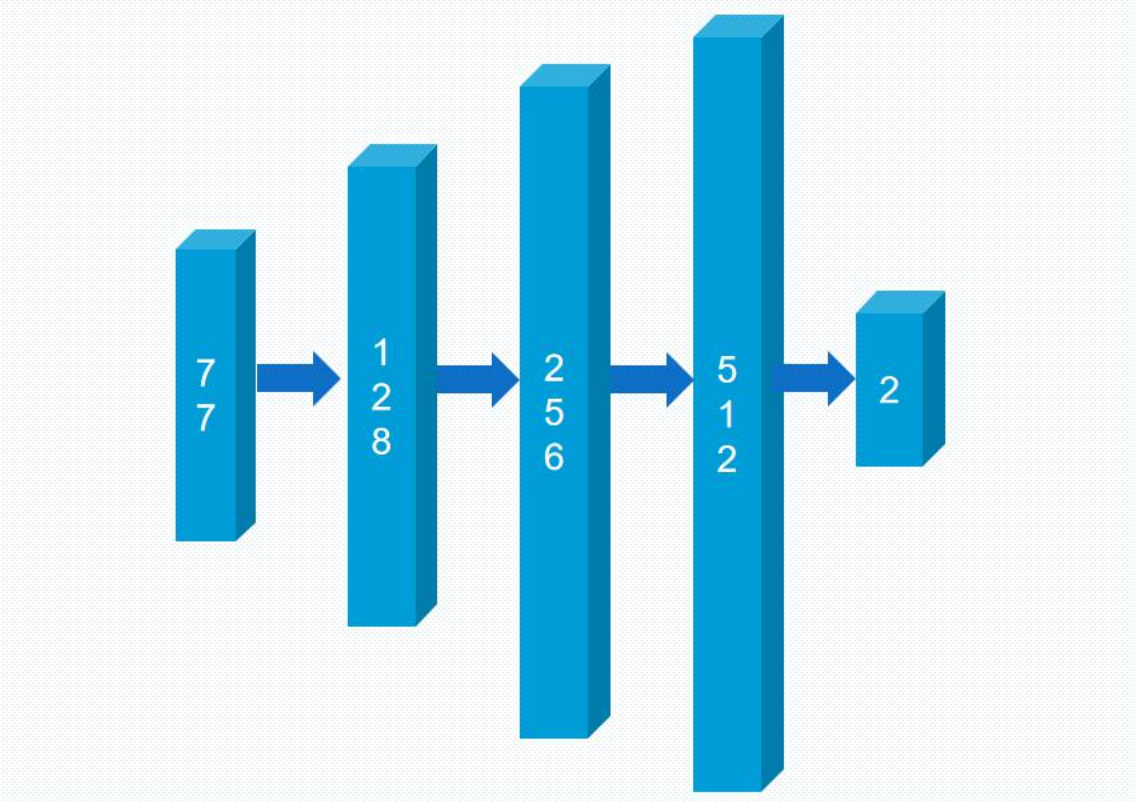
数据集是我们通过 github 中的开源代码中获得，其中包含 2011-2015 年部分印度公司的财务报表数据和欺诈标签。原始数据有 14464 行，共有 74 个属性特征，以及一个标签特征。

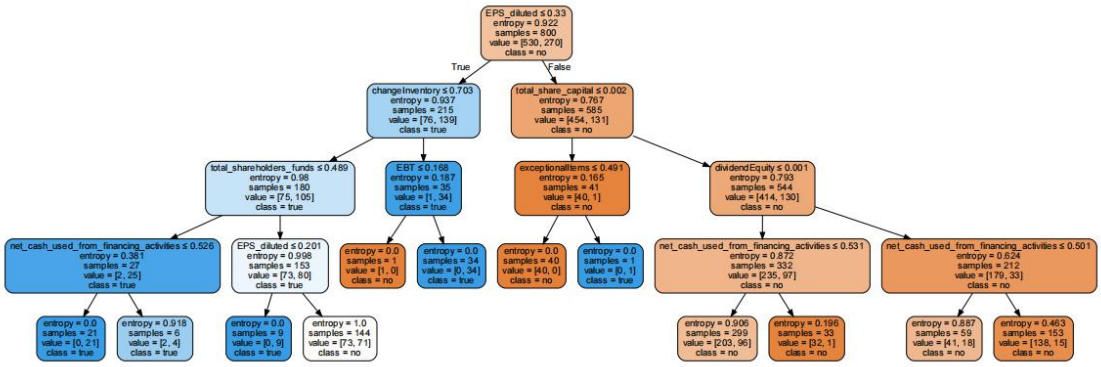
首先，在 14464 个数据中，欺诈数据只有 324 条只占总数据的 2.2%，直接用原始数据进行训练是不合理的，因为只预测为否的准确率就达到了 97%，所以我们需要对数据进行降采样。在这里，我们将所有的欺诈数据筛选出来，然后随机添加部分源数据组成 1000 条数据，作为最后的数据集。

因为，数据中包含字符类别特征，因此我们首先将这些类别特征进行整数化，然后用[0,1]代表替换分类标签，之后我们对数据进行归一化即可，之后我们就可以放到模型中进行训练了。

（3）模型的训练和预测

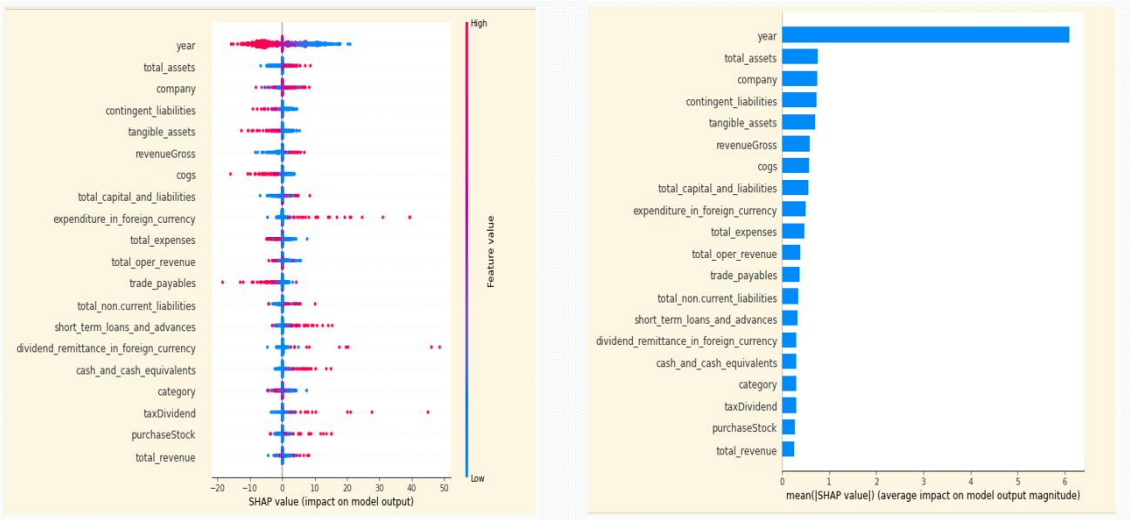
在这里，我们选用了神经网络模型、决策树、梯度提升树、随机森林、极度随机森林的方法，并进行了一些简单调参。下面是我们神经网络的模型架构和决策树。



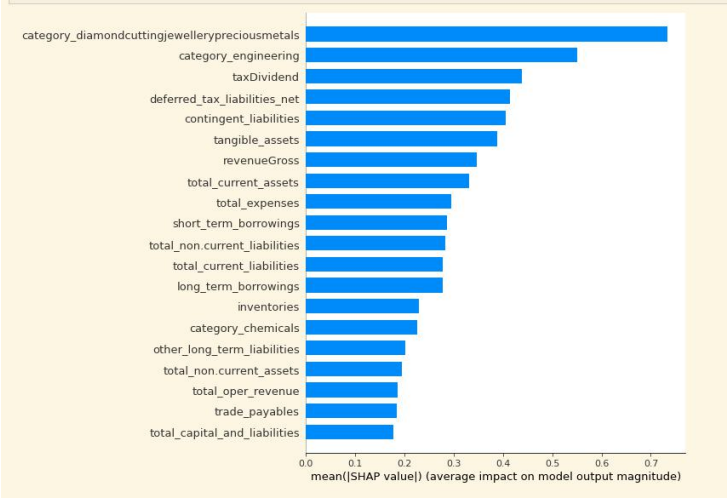


（4）结果的分析和再训练

在课上，我们讲过使用 shap 值对模型进行分析的方法，因此我们尝试使用这种方法对我们模型进行分析，观察不同特征的影响。



最后，我们发现 year 的影响最大，然后我们对数据进行分析发现，所有的欺诈数据中有 250 多例来自于 2015 年，所以我们 year 的这个特征可能会有些问题，之后我们会进行探讨，然后我们删除 year 进行再训练。



三、Year 特征的探讨

对于 year 特征，我们猜测一共有两种情况，分别为：2015 年印度的经济不景气，印度公司迫不得已进行欺诈以维持生存；或者数据库的作者在统计的过程中，没有找到 2015 之前的欺诈数据信息导致数据出现不均衡。

对此，我们尝试查找 2011-2015 年印度的 GDP 等反映印度公司或社会状况的信息，遗憾的是我们没有找到印度人们对当时社会的情感评论，只找到了当时的印度 GDP 的信息，以及当时税收、事业率等信息，以下是印度 GDP 的信息。

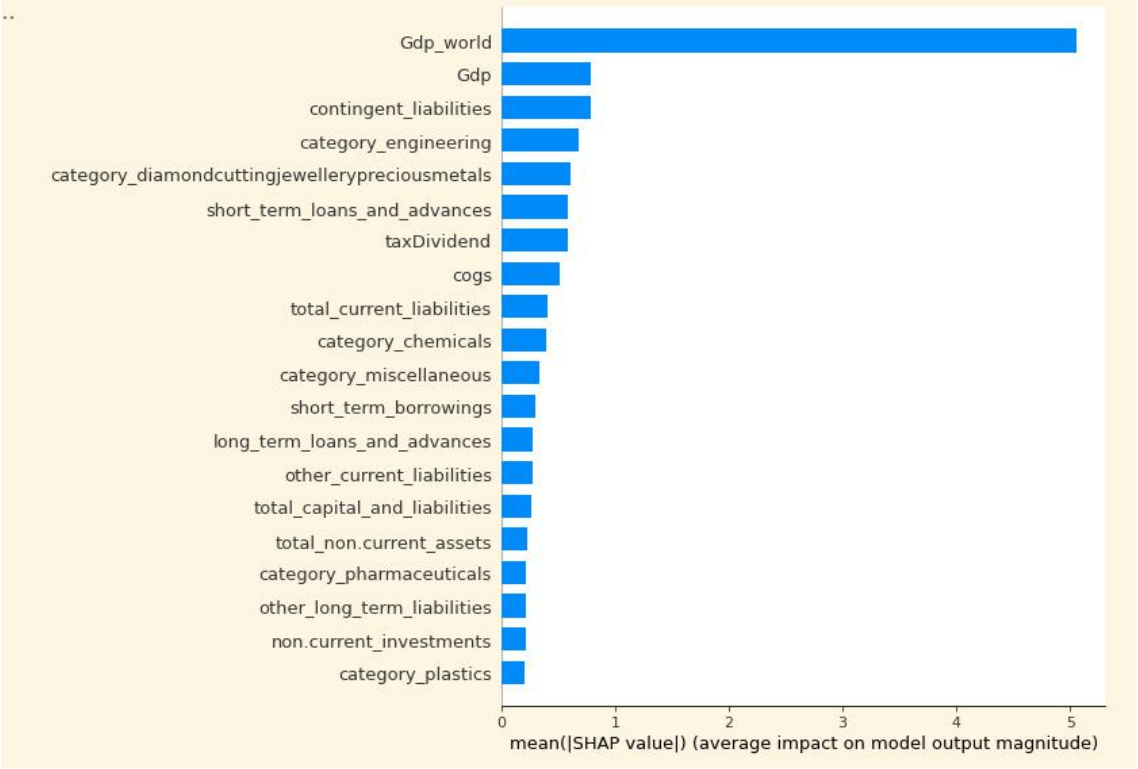
GDP 增长率：[5.24,5.46,6.39,7.41,8.00]；

GDP 占世界比例：[2.481,2.4312,2.401,2.5659,2.7961]；

以上信息均来自世界银行的统计结果，我们可以发现这些信息均不能反映印度在 2015 年的经济不好，因此我们认为造成这个情况的主要原因是数据库作者收集数据的不均衡所导致的。

在探索本部分的时候，我们尝试将以上数据作为特征加入到我们模型之中，结果发现其影响与 year 相同，因为以上数据是和 year 一一对应的，没有其他的意义。

如图，Gdp_world 代表 GDP 在世界 GDP 中的占比，Gdp 代表相比去年得增长率。

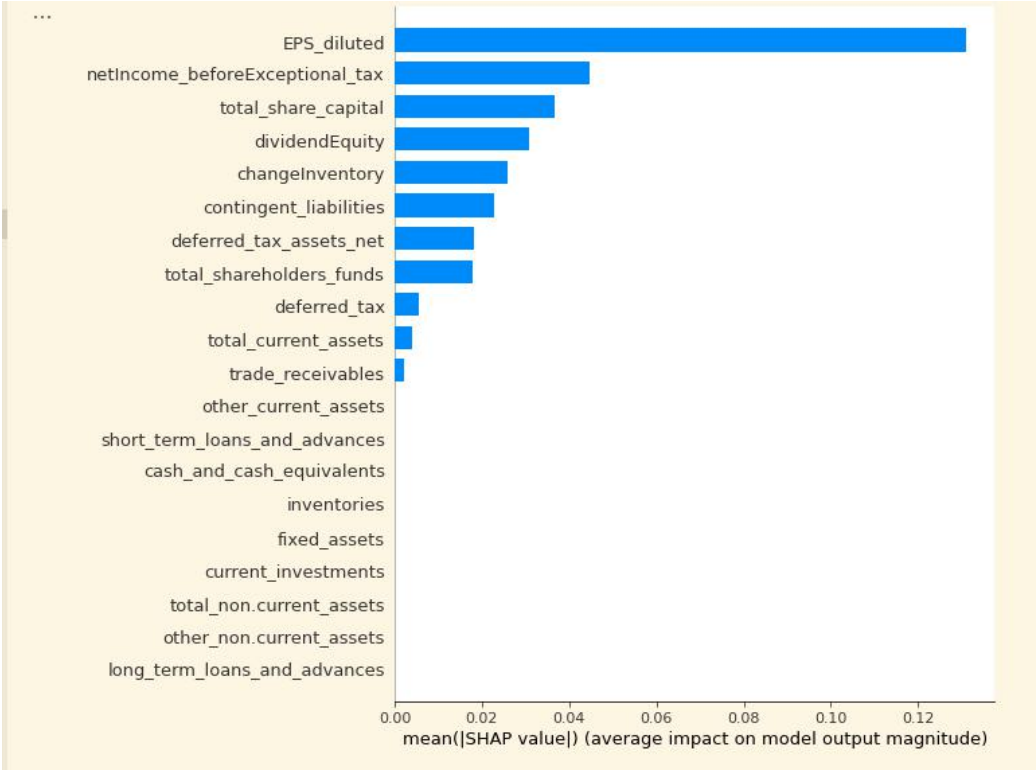


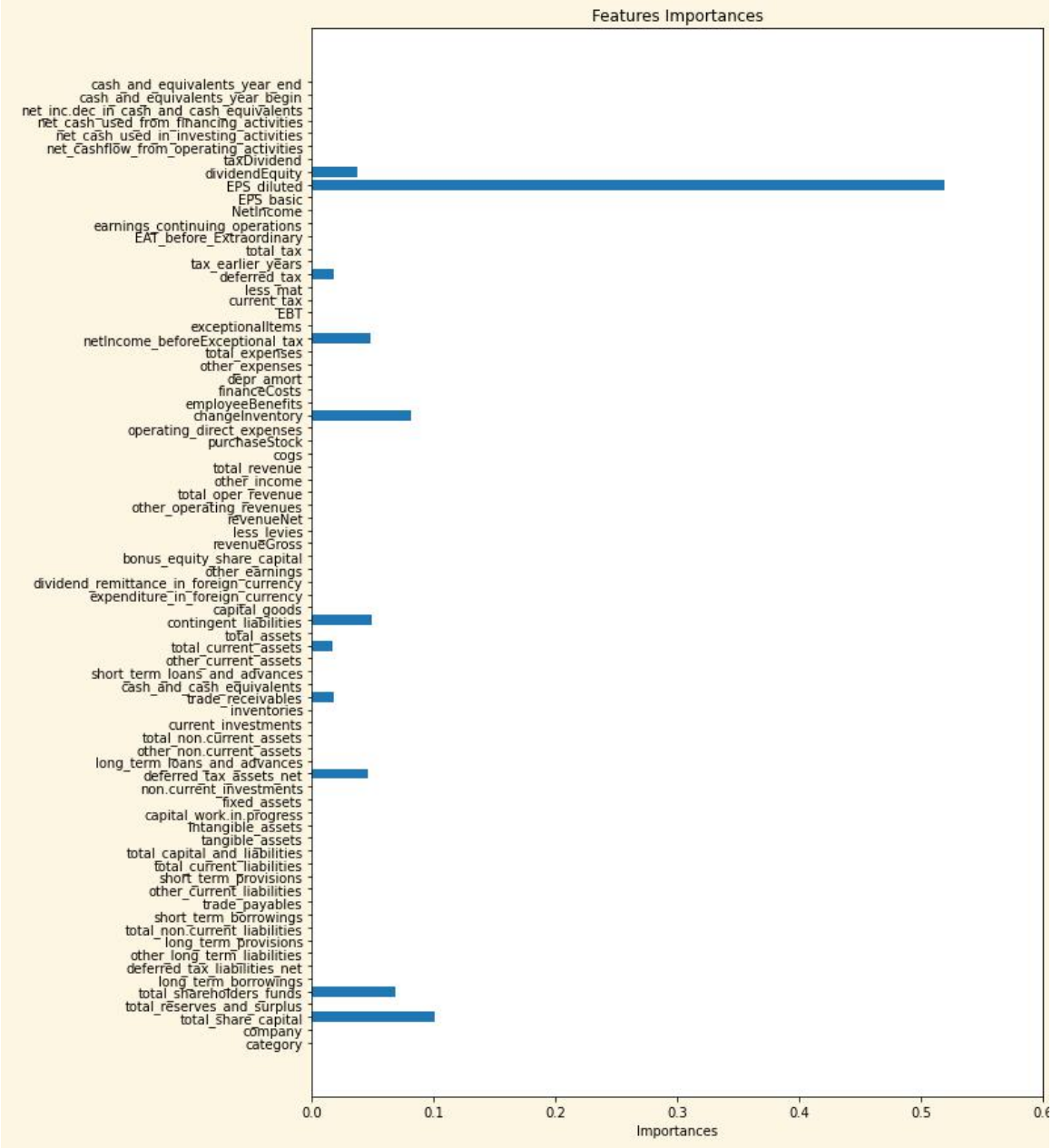
四、实验

在这一部分，我们给出我们使用不同的机器学习作出的不同结果，并进行对比和分析，如下表所示，我们发现 ExtraTree 准确度最好。

实验结果（去除 year 属性）					
算法	Accuracy	TP	TN	FP	FN
NN	0.715	33	110	22	35
DT	0.720	15	128	4	53
RF	0.765	35	118	14	33
ExtraTree	0.785	38	119	13	30
GBT	0.73	32	114	18	36

同时，我们也尝试使用决策树进行特征影响的分析，下面是不同库给出的分析结果，发现两者基本一致，但与 NN 的结果相比，具有较大的不同，因此我们可以根据不同模型给出的结果对我们的特征进行进一步的筛选，从而进一步优化我们的模型。





五、结语

在文中，我们使用不同的模型对数据进行了预测，通过对比我们发现 ExtraTree 的效果比其他几种方案效果更好，因此我们之后可以对 ExtraTree 进行进一步调优，优化准确度。

同时，在文中我们通过实验也发现数据中存在的一些问题，如欺诈数据主要集中在 2015 年导致数据进一步不平衡，削弱机器学习的能力。如果能有大量数据进行训练，我感觉准确率应该会更高。

不仅如此，我们认为机器学习在财务报表欺诈的方面会有较好的检测能力，那么通过机器学习对财务报表进行造假，从而进行欺诈的方法应该也具有较好的

可行性，那么这必然会进一步增加财务报表欺诈检测的难度，这不可避免是一个相互博弈的过程，检测和反检测的能力会逐渐升级。

目前，这个方面主要的遇到的问题，则是数据集的欠缺，毕竟进行财务报表欺诈的公司没有很多，或者说被发现的没有很多，因为一旦被公开则会遭到法律的制裁，所以数据量比较小，如本文中的数据集只有 1000，效果比较差。

我相信，随着各方的相互博弈，使用机器学习方法进行财务报表欺诈的检测是一个必然的趋势。

姓名：徐慧聪
学号：2019202363
时间：2022/1/13