# 上市公司财务报表错报的预测
# 基于中国资本市场 2002 年至 2017 年数据的分析

北京大学经济学双学位

潘斌

W15194142

2018 年 6 月 28 日

摘要

本文基于 2002 年至 2017 年中国资本市场的数据，比较 logit 模型和决策树类模型对于上市公司财务报表错报的预测力。本文首先对原始样本使用 SMOTE 算法进行过采样，以处理原始样本中类不平衡的问题。之后，本文进行 logit 模型、CART 和随机森林的算法的预测与评估。研究结果显示：SMOTE 算法确实能够提升三个模型查准率；传统的 logit 模型预测的精准度为 67.3%，而决策树类的模型精准度超过 80% 且具有更大的 AUC 值。在使用梯度优化法对于随机森林模型调参后，模型 AUC 达 0. 763。本文认为使用随机森林算法能够更好帮助利益相关者发现财务报表重大错报

关键词：财务报表错报　决策树　随机森林

**Abstract**

Based on the data of China's capital market from 2011 to 2017, this paper uses three statistical learning models to predict the misstatement of listed companies' financial statements. Before making the prediction, this paper first uses the SMOTE algorithm to oversample the original sample to deal with the unbalance problem in the original sample. Predict and evaluate logit models, CART, and Random Forest algorithms. The research results show that the problem of class imbalance does exist in the forecasting problem of misstatement of financial statements, and the SMOTE algorithm can indeed improve the precision of the three models; the accuracy of the traditional logit model in predicting the Chinese capital market is 67.3%, and the model accuracy of the decision tree and random forest is more than 80% with a larger AUC value. Afterwards, this paper uses the gradient optimization method to tune the random forest model. The AUC of random forest model after tuning is 0.763.

**Keywords:** Misstatements CART Random Forest

# 1 引言

我国资本市场乱象横生，形形色色的问题一经爆出，就会使得投资者、监管者和其他利益相关方蒙受巨大的损失。每年证监会都会披露上市公司的违规信息[1]。违规中有一部分是上市公司高管为了一己私利的个人违规行为，而另一部分是上市公司为了避免退市而进行的报表操纵。无论哪一类违规都会对公司的财务报表造成扭曲，影响公司价值。然而，这些违规证监会需要经过一段时间的调查才能得出结论，因此信息公开的时间都是滞后的（之后几个月甚至数年）。因此，本文意在建立一个能通过公司财务与非财务特征，快速评估公司当年是否可能违规的预测模型。同时，本文还想探究决策树等统计学习手段是否比传统 logit 回归在预测问题上更具有效率。

为什么有了审计师和审计报告还需要统计学习呢？被审计的上市公司需要付钱给审计师事务所。所以审计师和被审计单位是有利益相关关系的。在我国，每年出具的大部分审计报告都是标准无保留意见。然而，每年又有 10% 以上的公司因为错报被证监会审查。统计学习的介入，既能帮助审计师快速评估一个公司的情况，减少审计成本；又能给其他利益相关者以参考。

要建立预测模型主要有三个挑战：第一、检查违规行为是一个大海捞针的行为 [16]。每年违规的公司相对于没有违规（没有被发现违规）的公司是比例较少的。本文所用的数据集中，违规的比例是 16%[2]。数据的稀缺性[3]可能使得模型的拟合困难。第二、相对于较少的公司，有太多的变量可供选择 [16]。这些变量可能造成过拟合。本文所采用的变量是在经过国外检验的公认的对于识别有效的变量。第三、违规行为的异质性。违规行为如虚列资产和延迟披露等，其危害程度是不尽相同的。尽管如此，大部分文章将所有的违规视作相同类型进行处理[4]。这样做的好处是十分简便。本文也将沿用这种设定。

本文首先对训练集采用 SMOTE(Synthetic Minority Over-sampling Technique)[5] 算法进行过采样，以解决类不平衡的问题。之后，本文整理了盈余管理、公司绩效、非财务数据等特征，并进行如下几个模型的预测：第一是 logit 模型；第二是决策树；第三是随机森林。在进行时，本文使用交叉验证以避免过拟合。

本文发现：决策树和随机森林模型在经过调参后能够击败 logit 模型，具有更好的预测能力和更大的 AUC 值。进过调参后的随机森林模型的 AUC 值为 0.763，比 logit 模型高近 0.16。决策树模型的准确率为 83.1%，比 logit 回归高近 13 百分点。其次，本文给出了有决策树计算出的特征重要性排序，和连续变量的最优划分[5]。据此本文认为决策树和随机森林算法能够更好的帮助利益相关者尽快发现上市公司的报表问题。

本文的结构安排如下：第一部分是引言；第二部分回顾相关文献；第三部分阐述了本文的研究背景和初步分析；第四部分介绍本文使用的数据和统计学习模型；第五部分报告了实证的结果并进行模型评估；第六部讨论了本研究的结论和意义。

[1] 见证监会官网，行政处罚：http://www.csrc.gov.cn/pub/zjhpublic/

[2] 由于证监会调查滞后性，很多公司的违规行不会只持续一年，调查的结果可能是连续数年的违规。

[3] 即，类不平衡的问题 (imbalanced dataset)。

[4] 即，设定违规与不违规为一个二分变量，不对违规程度做区分。

[5] 由 CART 算法给出

# 2 背景以及相关分析

美国从本世纪初就开始关注财务报表的欺诈侦测 (fraud detection)，亦可称为财务报表错报的预测问题。在 2000 年，Bell 和 Carcello[2] 整理了一个包含 77 个错报公司和 305 个非错报公司的样本，并使用 logit 模型进行回归。他们发现较弱的内部控制、较快的公司增长速度和不正常的利润率是上市公司可能发生报表欺诈的标志。至此之后，一部分的文献集中于盈余管理与内部控制关系，这些文章将盈余管理当作财务报表质量的代理变量 [1][9][17]。另一部分则集中研究什么变量能够显著的帮助审计师解释财报质量 [8][12][10][6][3] 国内的研究集中在盈余管理与重大错报风险的相关关系上，一般认为盈余管理程度高的公司具有更高的重大错报风险 [21]。本文借鉴了上述研究所用的上市公司特征。

Varian[18] 认为当下的计量经济学学家应该更多借助统计学习工具去研究课题。近年来，很多的研究采用了如支持向量机 (SVM)、决策树分类器和一些集成算法等统计学习工具，并取得了比传统的计量工具更好的预测准确率 [19]。David 和 Hansen 等人 [19] 利用 Rule Ensemble (集成算法)，得到了一个 AUC 达 0.882 的预测模型。在 2015 年，Perols[16] 使用支持向量机 (SVM) 得到了关于错报公司和非错报公司的一个最大分割。

目前国内对上市公司违规的预测这类问题还鲜有人关注。因此，本文的利用较为时鲜的统计学习方法进行预测是具有一定价值的。同时，本文所用的方法是十分简单而直接的，易于操作，能够辅助决策者。

## 2.1 样本的不平衡问题

数据不平衡[6] 是各类欺诈预测模型需要首先解决的一个问题。非平衡的训练集会导致拟合无效[7]。关于样本不平衡的问题总结有两种解决方案：过采样 (Oversampling) 和欠采样 (Undersampling)。欠采样对训练集里面样本数量较多的类别（多数类）进行欠采样，抛弃一些样本来缓解类不平衡。但是这样做无疑会损失一些信息。而过采样是对训练集里面样本数量较少的类别（少数类）进行过采样，合成新的样本来缓解类不平衡。然而，直接重新抽样再加入样本并没有带来很大的性能提升 [14]，而且会造成过拟合 [5][8]。本文采用的 SMOTE[5] 是一种过采样算法。该算法流程如下：第一、对于少数类中每一个样本 $x$，以欧氏距离为标准计算它到少数类样本集中所有样本的距离，得到其 $k$ 近邻。第二、根据样本不平衡比例设置一个采样比例以确定采样倍率 $n$，对于每一个少数类样本 $x_n$，从其 $k$ 近邻中随机选择若干个样本，假设选择的近邻为 $x_n$。第三、对于每一个随机选出的近邻 $x_n$，分别与原样本按照公式 1 构建新的样本。具体的伪代

[6] 一般认为，如果若进行处理后，模型在验证集上的表现明显好于处理前，那么数据就是不平衡的

[7] 一个极端的例子：100 个个体的训练集中，正类样本 99 个，负类样本 1 个。训练过程中在某次迭代结束后，模型把所有的样本都分为正类，虽然分错了这个负类，但是所带来的损失实在微不足道，因为此时的 accuracy 已经是 99.9%

[8] 原文：If we replicate the minority class, the decision region for the minority class becomes very specific and will cause new splits in the decision tree. This will lead to more terminal nodes (leaves)as the learning algorithm tries to learn more and more specific regions of the minority class; in essence, overfitting. Replication of the minority class does not cause its decision boundary to spread into the majority class region.

码见附录6.3这种算法能提高分类器的 AUC 值，改善分类器性能。这种算法可以通过 python 中的 imblearn 中的 over_sampling 包轻松实现。

$$x_{new} = x + rand(0,1) * (\widetilde{x} - x) \tag{1}$$

## 2.2 统计学习算法

对于每个模型，本文都使用交叉验证法 (k-fold cross-validation)[9]。首先本文将样本分成五份，随机去一份作为验证集，剩下作为训练集。其次，对 5 个训练集做 SMOTE 算法，改善训练集的平衡性。之后，本文使用三个模型分别进行训练、调参。对于随机森林模型，本文使用梯度优化法[10]进行调参。最后，本文对模型进行多个指标的评估。由于 Logit 回归最为常见，本文在此就不做过多赘述。

### 2.2.1 分类回归树 (CART,Classification And Regression Tree)

本文使用的 CART(Classification And Regression Tree) 算法最先由 Breiman[15] 提出。CART 假设决策树是二叉树，内部结点特征的取值为“是”和“否”，左分支是取值为“是”的分支，右分支是取值为“否”的分支。这样的决策树等价于递归地二分每个特征，将输入空间即特征空间划分为有限个单元，并在这些单元上确定预测的概率分布，也就是在输入给定的条件下输出的条件概率分布。其算法的核心是根据输入特征给出一个最佳的分组，再从每个分组的众多取值中获取一个最佳的分割。选择的标准是 Gini 系数。即：

$$Gini(D_j) = 1 - \sum_{t=0}^{c-1} p_i^2 \tag{2}$$

其中，c 是数据集 $D_j$ 中决策类的个数，$p_i$ 是第 i 个决策类在 D 中的比例。Gini 系数表示从相同的总体中随机抽取两个样本后在，这两个样本来自不同类别的概率。之后，再将数据集划分成多个数据子集，这些数据子集划分前后的 Gini 系数与划分前的 Gini 系数加权和的差为：

$$G(A) = Gini(D) - \sum_{j=1}^{k} \frac{|D_j|}{|D|} Gini(D_j) \tag{3}$$

其中，A 是侯选属性，k 是该属性的分支个数，D 是未使用 A 进行划分的数据集；$D_j$ 是由 A 划分而成的子数据集。在所有属性中具有最大 $G(A)$ 的属性即选为当前划分的方式。待决策树生成后，CART 算法用验证数据集对已生成的树进行剪枝[11]并选择最优子树，这时损失函数最小作为剪枝的标准。本文选择 CART 算法的原因是该算法能返回一个每个连续变量的最优分割，这个分割能一定程度上作为公司错报的标志 (a red flag)。

[9] 本文的代码和数据公开：
[10] python 中的 GridSearchCV 能够自动返回最优参数。
[11] 分的过于细致的决策树，非常有可能造成过拟合

### 2.2.2 随机森林 (Random Forest)

随机森林是一个包含多个决策树的分类器，并且其输出的类别是由个别树输出的类别的众数而定。该算法仍然是由 Breiman[4] 在这 2001 提出的一种集成学习 (Ensemble Learning) 的 bagging 算法。随机森林与 Bagging 的简单过程见附录6.4。该算法在精准度上具有一定优势，而且能返回每个变量的重要性程度，对于财务报表欺诈的预测而言具有一定意义[12] 同时，这类算法对于缺失值和极端值不敏感。

# 3 数据与变量

## 3.1 数据来源

本文所采用的数据是国泰安 (CSMAR) 数据库的子库上市公司违规数据库。违规处理数据库收集了 1994 以来在上海证券交易所和深圳证券交易所上市的有违规行为[13]的上市公司公布的企业公告，证监会指定媒体的报道及监管机构所出的公告等相关数据。由于其他特征信息的不足，本文选择 2002 年至 2017 年的数据。经过梳理，本文总结出的有效违规信息 5639 条（同一上市公司可能在同一年份涉及多项违规），涉及的违规种类见表1。本文关于上市公司的财务指标、公司属性、公司市值等特征也来自国泰安数据库。

表 1: 2002 年至 2017 年上市公司违规种类汇总表

| 类别 | TypeID | Freq | Percent | Cum |
|---|---|---|---|---|
| 虚构利润 | P2501 | 230 | 2.44% | 2.44% |
| 虚列资产 | P2502 | 39 | 0.41% | 2.85% |
| 误导性陈述 | P2503 | 969 | 10.27% | 13.12% |
| 推迟披露 | P2504 | 1,952 | 20.68% | 33.80% |
| 重大遗漏 | P2505 | 1415 | 14.99% | 48.79% |
| 披露不实 | P2506 | 333 | 3.53% | 52.31% |
| 欺诈上市 | P2507 | 8 | 0.08% | 52.40% |
| 出资违规 | P2508 | 5 | 0.05% | 52.45% |
| 擅自改变资金用途 | P2509 | 104 | 1.10% | 53.55% |
| 占用公司资产 | P2510 | 271 | 2.87% | 56.43% |
| 内幕交易 | P2511 | 140 | 1.48% | 57.91% |
| 违规买卖股票 | P2512 | 919 | 9.74% | 67.64% |
| 操纵股价 | P2513 | 20 | 0.21% | 67.86% |
| 违规担保 | P2514 | 166 | 1.76% | 69.62% |
| 一般会计处理不当 | P2515 | 458 | 4.85% | 74.47% |
| 其他 | P2599 | 2410 | 25.53% | 100.00% |
| | 合计 | 9439 | | |

本文对于违规的行业和年份特点进行总结得到表2。原始样本中平均违规的比例

[12] 比如公司的市值在随机森林的预测中很重要，那么很可能审计师和监管者们应该更关注上市公司高管违规股票买卖套现行为。

[13] 违规类型编码对应内容：P2501= 虚构利润; P2502= 虚列资产; P2503= 虚假记载 (误导性陈述);P2504= 推迟披露; P2505= 重大遗漏; P2506= 披露不实 (其它); P2507= 欺诈上市; P2508= 出资违规; P2509= 擅自改变资金用途; P2510= 占用公司资产; P2511= 内幕交易; P2512= 违规买卖股票; P2513= 操纵股价; P2514= 违规担保; P2515= 一般会计处理不当; P2599= 其他

16%, 为计算机通信行业、化工行业、医疗制造业和房地产是违规比例相对比较高的行业。本文在后续的分析中，将行业分成了第一产业、第二产业、第三产业（排除金融业）和金融业[14]。第一产业的违规比例最高，有 27.9%；金融业其次，上市公司违规的比例为 19.4%。上市公司违规比例在时间序列上没有比较明显的规律。2012 年是违规比例最高的一年，高达 23.12%。

表 2: 不同年份、分行业违规情况表

| 违规行业 | | | | 每年违规比例 | | | |
|---|---|---|---|---|---|---|---|
| 行业 | 无违规 | 违规 | 占比 | 年份 | 无违规 | 违规 | 比例 |
| 计算机、通信和其他 | 2096 | 404 | 7.90% | 2002 | 1032 | 184 | 15.13% |
| 化学原料及化学制品 | 1602 | 353 | 6.90% | 2003 | 1118 | 163 | 12.72% |
| 医药制造业 | 1589 | 311 | 6.08% | 2004 | 1209 | 162 | 11.82% |
| 房地产业 | 1592 | 310 | 6.06% | 2005 | 1232 | 127 | 9.35% |
| 电气机械及器材制造业 | 1470 | 301 | 5.89% | 2006 | 1249 | 138 | 9.95% |
| 软件和信息技术服务业 | 988 | 190 | 3.72% | 2007 | 1260 | 201 | 13.76% |
| 批发业 | 797 | 186 | 3.64% | 2008 | 1331 | 250 | 15.81% |
| 专用设备制造业 | 1184 | 173 | 3.38% | 2009 | 1379 | 299 | 17.82% |
| 通用设备制造业 | 867 | 161 | 3.15% | 2010 | 1684 | 326 | 16.22% |
| 汽车制造业 | 811 | 152 | 2.97% | 2011 | 1806 | 479 | 20.96% |
| 零售业 | 902 | 138 | 2.70% | 2012 | 1882 | 566 | 23.12% |
| 非金属矿物制品业 | 681 | 130 | 2.54% | 2013 | 1900 | 527 | 21.71% |
| 有色金属冶炼及 | 583 | 120 | 2.35% | 2014 | 2012 | 460 | 18.61% |
| 电力、热力生产和供电 | 866 | 117 | 2.29% | 2015 | 2121 | 530 | 19.99% |
| 土木工程建筑业 | 534 | 110 | 2.15% | 2016 | 2502 | 410 | 14.08% |
| 互联网和相关服务 | 336 | 102 | 1.99% | 2017 | 3047 | 290 | 8.69% |
| 农副食品加工业 | 327 | 102 | 1.99% | 合计 | 26764 | 5112 | 16.04% |

注：行业划分根据 2012 年证监会制定的行业划分；完整的表格见附录

## 3.2 变量定义

本文将上市公司的财务特征分成以下几类以方便进行统计学习。变量的选择和分类是仿照 Dechow[8]2011 年的文章所设计的。同时，本文还加入了一些中国特色的变量：比如上市公司是否是国有企业，上市公司的总经理和董事长是否兼任等。详细的变量定义表见3

**盈余管理** 一般来说，盈余管理 (Earning Management) 就是企业管理当局在遵循会计准则的基础上，通过对企业对外报告的会计收益信息进行控制或调整，以达到主体自身利益最大化的行为。关于盈余管理的计量，主流的方法是 Dechow[7] 在 1995 年提出的修正的 Jones 模型[15]。尽管盈余管理本身不违法，这种行为是对真实信息的一种歪曲。上市公司在游走于法律的行为可能由盈余管理的数值来反映 [20]。

**公司绩效** 公司异常的销售增长或者净利润增长可能来自虚构交易等违规行为 [8]。因此，本文引入多个评价公司绩效的特征。同时，由于中国证监会规定上市连续三年亏损

[14] 银行等金融机构的资产负债表显著不同于其他公司，故将其单列一类，以示区分。
[15] 关于此方法的计算详见附录6.1

表 3: 变量定义表

| 特征 | 中文 | 定义 |
|---|---|---|
| lemon | 违规 | 本文研究的对象；本文简单的将其分为违规和不违规两类 |
| accurals | 总应计利润 | 营运资本的变动-现金变动 + 短期借款变动-当期折旧 |
| da | 盈余管理水平 | 利用 Jones 模型得到的衡量企业盈余管理水平的指标 |
| soft_asset | 软资产 | 总资产剔除固定资产后的份额 |
| ch_sales | 销售收入变得率 | 本年销售收入与上年销售收入的比值-1 |
| ch_assets | 总资产变动率 | 本年总资产与上年总资产的比值-1 |
| ch_rev | 应收账款变动率 | 本年应收账款与上年应收账款的比值-1 |
| ch_inv | 存货变动率 | 本年存货与上年存货的比值-1 |
| ch_cs | 现金销售 | 销售收入变动中扣除应收账款变化的部分 |
| ch_roa | 总资产收益率变动 | 本年 ROA 与上年 ROA 的比值-1 |
| ch_emp | 人员流动率 | 本年员工数量与上年员工数量的比值-1 |
| llemon | 上年是否违规 | 标识上市公司上一年度是否违规 |
| stateown | 国企 | 标识上市公司是否为国企 |
| pe | 市盈率 | 当年股价/美股收益 |
| marketvalue | 市值 | 上市公司市值 |
| tobinq | 托宾 Q | 上市公司市值与其重置成本的比值 |
| booktomarket | 市价对帐面价值比率 | 上市公司账面价值与市值的比值 |
| lev | 资产负债率 | 负债与资产的比值 |

有退市可能，本文在公司特征中加入了公司是否连续两年亏损这一特征[16]。资产负债率能够反映企业的现在资本结构。过高的资产负债率预示着企业的偿债能力不强，抗风险能力弱。

**非财务指标** 公司的人员变动会反映上市公司的潜在的问题。一方面，当年财政年度一场的雇佣员工数量变化可能是公司出现问题的一个信号 [8]。另一方面，上市公司高管的兼任情况（比如，总经理和董事长兼任）可能反映公司较弱的内控 [21]。

**市场相对指标** 上市公司违规中有一类违规是高管通过内幕交易操纵公司股票，从中攫取暴利。因此，本文选择了一些市场指标特征，如 PE、托宾 Q 等，来反应上市公司的市值特征。

## 3.3 描述性统计

本文对于上述变量进行了描述性统计4。原始样本中，出现财务报表错报的公司占 16%; 国有企业占 45.46%; 连续出现错报的上市公司占 16.7%。与没有出现错报的上市公司相比6.2，出现错报的上市公司在均值意义上，具有更高的销售增长率; 较高的资产变动率；较低的固定资产变动率; 较低的现金销售水平；跟高的资产负债率以及跟高的人员流动率。假若一个上市公司基于某种目的粉饰报表，实物[17]与现金是比较难以操控的；而销售额[18]和总资产水平是更容易被操控的。同时，本文还发现，出现错报的上市公司在当年的市值指标都显著低于市场中间水平（经过标准化后）。然而大部分错报都

[16]信息来自：中国证券监督管理委员会关于《亏损上市公司暂停上市和终止上市实施办法（修订）》的通知http://www.csrc.gov.cn/pub/newsite/ssb/ssflfg/bmgzjwj/ybgd/200911/t20091110_167679.htm

[17]固定资产和现金需要多个账目登记，这些账目是有不同的人来掌管的，若想造假，所有部门必须齐心协力、沆瀣一气。现金同理

[18]虚拟交易、同时需开发票能够虚增本年度的销售额

不是本年对被发现的。这里我们可以理解为有内部人士走漏了风声（有效市场假说），使得上市公司在报表上的粉饰对于资本市场的是无效的。

表 4: 描述性统计

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| lemon | 31,876 | 0.160371 | 0.366956 | 0 | 1 |
| accurals | 29,555 | -3.67E+08 | 6.91E+09 | -4.61E+11 | 3.30E+11 |
| wc_acc | 25,185 | 0.000888 | 0.129186 | -2.33617 | 3.446984 |
| da | 28,031 | -3.12E+08 | 6.91E+09 | -4.61E+11 | 3.30E+11 |
| soft_asset | 31,876 | 0.75637 | 0.180694 | 0.029079 | 1.206255 |
| ch_sales | 30,764 | -6.34042 | 1102.192 | -191777 | 8072.186 |
| ch_assets | 29,871 | 0.711519 | 36.05459 | -0.99972 | 4719.612 |
| ch_rev | 28,612 | 0.012328 | 0.071407 | -1.91049 | 1.119883 |
| ch_inv | 28,403 | 0.02187 | 0.086035 | -1.60021 | 1.549514 |
| ch_cs | 28,321 | 5.93E+09 | 5.23E+10 | -9.70E+11 | 2.89E+12 |
| ch_roa | 28,689 | 0.00528 | 0.315686 | -32.1618 | 20.95762 |
| ch_emp | 28,677 | 1.06917 | 2.165678 | 0.000785 | 154 |
| llemon | 28,785 | 0.167275 | 0.373227 | 0 | 1 |
| stateown | 31,876 | 0.454637 | 0.497946 | 0 | 1 |
| pe | 31,866 | 96.37426 | 2751.74 | -129431 | 420284.6 |
| marketvalue | 31,876 | 1.35E+10 | 7.77E+10 | 1.98E+07 | 5.67E+12 |
| tobinq | 31,876 | 4.095862 | 285.9928 | 0.006954 | 50939.54 |
| booktomarket | 31,876 | 1.042275 | 1.684249 | 0.00002 | 143.8035 |
| lev | 31,876 | 0.530446 | 5.084022 | -0.1947 | 877.2559 |

# 4 统计学习模型

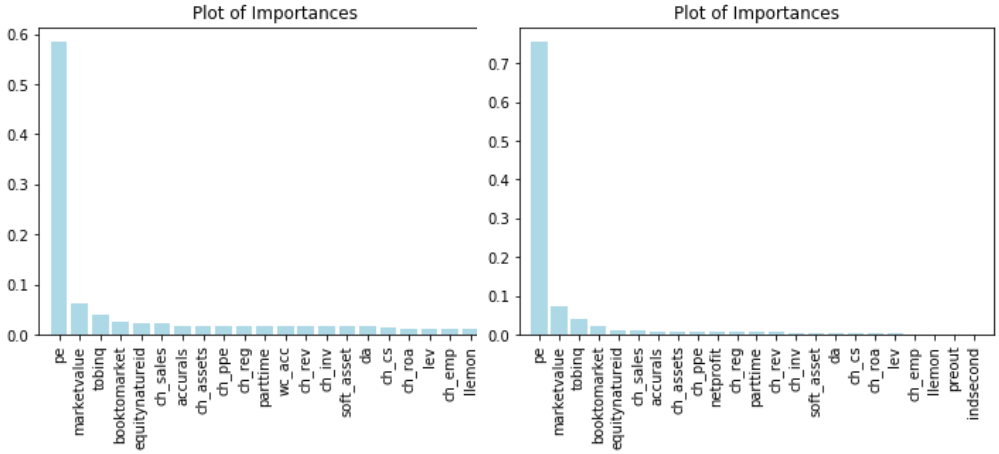
## 4.1 模型的建立

本文通过 python 编写了关于该问题的一个类。读入数据后，本文先将 Y 变量和特征变量分开，再使用交叉验证的办法得到 5 个训练集和 5 个测试集。之后，本文对训练集使用 SMOTE 算法，进行过采样，得到新的训练集。训练集会通过 logit 回归, CART 以及随机森林三个算法进行训练。通过测试集可以进行模型评估、调参。程序会输出特征重要性排序，混淆矩阵，AUC 值和学习曲线等指标。本文将进行以下几个模型的拟合：1）Logit 模型；2）过采样 +Logit 模型；3）CART；4）过采样 +CART；5）随机森林算法；6）过采样 + 随机森林算法；7）经过梯度法调参的随机森林算法。

## 4.2 模型的结果

决策树和随机森林都能够返回特征对于模型预测的重要性排序。本文发现对于中国的资本市场而言，PE（市盈率）是一个很重要的指标。根据描述性统计，原始样本中出现上市公司出现错报的年份市盈率会更低。在本文的框架下，市盈率的异常与否能够明显的分开测试集中发生财报错报的公司和没有发生错报的公司。与变量不能显著解释这一点可能是由于本文特征选择不够具有代表性造成的。本文选择的变量是文献中对于美

国资本市场有效的特征。在有限的时间成本下，本文没有在去跟多的扩展特征。




(a) CART (b) Random Forest

图 1: 重要性排序

### 4.2.1 精准度 (Accuracy)

模型的精准度的是分类正确的个体再测试集中的比重。从表6中可以看出，logit 模型的预测精准度最低只有 67.3% [19]，这与 Dechow[6] 的结果比较接近。CART 算法的精准度是 82.1%。再经过最大深度选择后，决策树的精度达到 84.7%。这一结果超过了未调参的随机森林模型。后者经过调参后模型精度达到 85.5%。从精度这个角度，随机森林模型具有更好的预测力。

### 4.2.2 查准率 (Precision)、查全率 (Recall) 和 F1

模型的预测结果可以用混淆矩阵8表示。查准率的定义式 $P(Precision) = \frac{TP}{TP+FP}$，即测试集中预测为错报的公司中真正错报的比例;查准率的定义式 $R(Recall) = \frac{TP}{TP+FN}$，即测试集中所有错报的公司中被模型发现的比例;$F1$ 是 $P$ 与 $R$ 的调和平均数。传统的 logit 模型在查准率的表现上不如树一类的统计学习模型。同时，SMOTE 算法的确能够提高查准率，而使用旧的样本更会使模型把预测值归为非异常类。这也说明原始的样本确实存在类不平衡的问题。logit 模型的查准率为 13%，远低于 CART 和随机森林超过一半的查准率。实践中，如果利益相关者关心“错杀”的比例，那么他会更关注从查准率、查全率和 F1 的角度，随机森林算法具有更好的性能。

[19]若使用原始的样本，那么 logit 模型精准度将会只有 45.3%

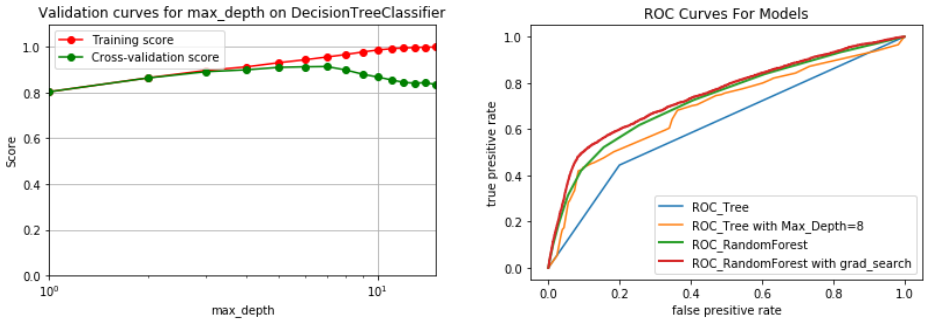
表 5: 查全率、查准率和 F1

| Model | precision | recall | F1-score |
|---|---|---|---|
| Logit | 0.132 | 0.871 | 0.229 |
| Logit without SMOTE | 0.121 | 0.845 | 0.212 |
| CART | 0.500 | 0.438 | 0.467 |
| CART without SMOTE | 0.368 | 0.541 | 0.438 |
| Random Forest | 0.545 | 0.477 | 0.509 |
| Random Forest without SMOTE | 0.349 | 0.574 | 0.434 |

## 4.3 模型的评估与调参

### 4.3.1 决策树的 Max Depth

在表2(a)中，本文展示了利用 validation curve 计算不同深度训练集和测试集交叉验证得分。如果不限制最大深度，就会发生过拟合的问题，即模型在训练集与测试集上的表现发生背离。对于中国资本市场侦测财报错报的问题，本文的结果是将决策树的深度参数限制为 8 为最优。




(a) Validation Curve for Trees (b) ROC Curve for Trees

图 2: 模型评估

### 4.3.2 AUC 与 ROC 曲线

评价分类器的一个主要指标是 ROC 曲线和 AUC 值。受试者工作特征曲线(Receiver Operating Characteristic Curve，简称 ROC 曲线)，又称为感受性曲线（Sensitivity Curve）。最早是由 Lusted[13] 在 1988 年提出的。ROC 曲线是以假阳性概率（False Positive Rate）为横轴，真阳性（True Positive Rate）为纵轴所组成的坐标图，和受试者在特定条件下由于采用不同的 thresholds 得出的不同结果画出的曲线。AUC（Area Under Curve）被定义为 ROC 曲线下的面积，AUC 值是一个概率值 [11]，即当随机挑选一个正样本以及一个负样本，当前的分类算法根据计算得到的 Score 值将这个正样本排在负样本前面的概率。AUC 值越大，当前的分类算法越有可能将正样本排在负样本前面，即能够更好的分类。

本文计算了三种模型基于不同设定下的 AUC[20]值。Logit 模型的 AUC 值最低。决策

[20] Fawcett(2006) 原文：The AUC value is equivalent to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example.

树和随机森林的结果比较接近，均超过了 0.7。同时，这一结果略弱于 David 和 Hansen 等人 [] 利用 Rule Ensemble 得到的 AUC。据此，本文认为对于财务报表错报侦测的问题，决策树分类器比 logit 模型根据有预测能力，能够辅助利益相关者进行决策。

表 6: 精准度与 AUC

| Model | Accuracy | AUC | std |
|---|---|---|---|
| Logit | 0.673 | 0.595 | 0.016 |
| CART | 0.821 | 0.693 | 0.026 |
| Tree with max_depth = 8 | 0.847 | 0.539 | 0.016 |
| Random Forest | 0.832 | 0.736 | 0.009 |
| Random Forest with tunning | 0.855 | 0.763 | 0.007 |
| Note: SMOTE is used before modeling | | | |

#### 4.3.3 随机森林调参

本文进一步对随机森林模型使用梯度优化模型进行调参。在 python 中可以直接调用 gridsearch 进行并行调参。调整的参数包括：1) 混乱都下降标准选择“基尼系数”还是“信息熵”;2) 最大深度；3) 单个模型最大特征数目;4) 迭代次数。具体操作见代码源。调参的结果为一个最大深度为 12、特征数位 5、迭代次数为 31 的使用'Entropy' 为标准的模型。调参使得随机森林模型预测能力提升：模型精确度上升 85.5%，模型 AUC 达到 0.763。

## 5 结论与反思

本文基于 2002 年至 2017 年中国资本市场的数据，通过三个统计学习模型来预测上市公司财务报表错报。在进行预测前，本文首先对原始样本使用 SMOTE 算法进行过采样，以处理原始样本中类不平衡的问题。之后本文使用 python 编写了一个进行预测的类，并进行 logit 模型、CART 和随机森林的算法的预测与评估。研究结果显示：类不平衡的问题确实存在与财务报表错报的预测问题之中，而 SMOTE 算法确实能够提升三个模型查准率；传统的 logit 模型在中国资本市场进行预测的精准度为 67.3%，而决策树类的模型精准度有 80% 以上而且具有更大的 AUC 值。这些结论告诉足以说明，本文的框架下，决策树类的统计学习模型对于财务报表的错报更具有预测力、性能更好。最后，本文使用梯度优化法来实现对于随机森林模型的自动调参，调参后的模型性能更佳。

本文意义是为利益相关这判别某公司是否具有财务报表错报提供只管的决策辅助。在实际使用中，利益相关者如对冲基金研究员会更多的在意某一家或者某一个具体行业的上市公司是否发生重大错报，而不是概率平均值。统计学习的模型提供了一个近乎于黑箱的解决思路，可能这样的思路会在实际应用中更直观。

此外，本文通过决策树类的模型得到了一个关于预测中国资本市场财报错报预测的重要性排序。在研究某一个公司是否发生财务报表错报时，优先关注重要性高的特征能

够帮助利益相关方更好的了解公司的报表是否造假，并为后续的研究提供一些信息。这些特征是否有效，还有赖于更多的研究中国资本市场财报错报的文章出现。

对于经济学模型来讲，统计学习的很多操作都看似“黑箱”一般，即只关注输入输出。这种模型与传统的结构方程和离散选择模型孰优孰劣不能一概而论。在过去，过度而随意的加变量，会被经济学家成为 Data Mining。统计学家也将统计学习看作一种下下策的出路。然而，统计学习确实可以帮助经济学家进行变量筛选 (LASSO、岭回归)，为经济学研究体重很多思路。统计学习中的抽样算法能帮助解决样本的很多问题。交叉验证的思路为模型的稳健性提供支持。目前统计学习（机器学习）正在谷歌等公司的带领下，触及因果推断。

本文有很多粗糙和不够完善的部分。首当其冲的是关于 Y 变量粗糙的选择。本文将所有的错报一视同仁，然而错报也是分危害程度的。若是将错报分成低级、中级、重大来进行学习，可能更具有实际意义。此外，本文将面板数据当作截面数据来使用，将不同年份的同一个体视作不同的个体，没有讲年份因素考虑进入上市公司特征。这样做简便，但是会损失预测力。还有一些做的不细致的地方，由于时间和精力没有做进一步探讨。

# 参考文献

[1] Mark S. Beasley. An empirical analysis of the relation between the board of director composition and financial statement fraud. *The Accounting Review*, 71(4):443–465, 1996.

[2] Timothy B. Bell and Joseph V. Carcello. A decision aid for assessing the likelihood of fraudulent financial reporting. *Auditing A Journal of Practice*, 19(1):169–184, 2000.

[3] Messod D. Beneish. Detecting gaap violation: implications for assessing earnings management among firms with extreme financial performance. *Journal of Accounting and Public Policy*, 16(16):271–309, 1997.

[4] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, 2002.

[6] Patricia M. Dechow, S. P. Kothari, and Ross L. Watts. The relation between earnings and cash flows. *Social Science Electronic Publishing*, 25(2):133–168, 2006.

[7] Patricia M. Dechow, Richard G. Sloan, and Amy P. Sweeney. Detecting earnings management. *The Accounting Review*, 70(2):193–225, 1995.

[8] Patricia M. Dechow, G. E. Weili, Chad R. Larson, and Richard G. Sloan. Predicting material accounting misstatements. *Contemporary Accounting Research*, 28(1):17-82, 2011.

[9] DechowP, SloanR, and SweeneyA. Causes and consequences of earnings manipulation: An analy. 1996.

[10] Michael Ettredge, Lili Sun, Picheng Lee, and Asokan Anandarajan. Is earnings fraud associated with high deferred tax and/or book minus tax levels? *Social Science Electronic Publishing*, 27(1):1–33, 2011.

[11] Tom Fawcett. Roc graphs with instance-varying costs. *Pattern Recognition Letters*, 27(8):882–891, 2006.

[12] Andrew J Felo. Using nonfinancial measures to assess fraud risk. 2007.

[13] D. J. Goodenough, K Rossmann, and L. B. Lusted. Radiographic applications of receiver operating characteristic (roc) curves. *Radiology*, 110(1):89, 1974.

[14] N. Japkowicz. The class imbalance problem: Significance and strategies, proceedings of the second international conference on artificial. 2000.

[15] Breiman LI, J.H. Friedman, RA Olshen, and C.J. Stone. Classification and regression trees (cart). *Encyclopedia of Ecology*, 40(3):582–588, 1984.

[16] Johan Perols, Robert M Bowen, Carsten Zimmermann, and Basamba Samba. Finding needles in a haystack: Using data analytics to improve fraud prediction. *Social Science Electronic Publishing*, 2015.

[17] Scott L. Summers and John T. Sweeney. Fraudulently misstated financial statements and insider trading: An empirical analysis. *Accounting Review*, 73(1):131–146, 1998.

[18] Hal R Varian. Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28(2):3–27, 2014.

[19] David G Whiting, James V Hansen, James B Mcdonald, Conan Albrecht, and W. Steve Albrecht. Machine learning methods for detecting patterns of management fraud. *Computational Intelligence*, 28(4):505–527, 2012.

[20] 周晓苏 and 周琦. 基于盈余管理动机的财务重述研究. 当代财经, (2):109–117, 2011.

[21] 孙欣然. 盈余管理、分析师盈余预测与重大错报风险. PhD thesis, 吉林大学, 2015.

# 6 附录

## 6.1 修正的 Jones 模型的计算

修正的 jones 模型在回归的过程中应当使用分行业分年份回归，其回归模型设定如下如下：

$$TA_{i,t} = \beta_1(\frac{1}{A+i,t-1}) + \beta_2(\Delta_{i,t} - \Delta_{i,t} + \beta_3 PPE_{i,t} + \varepsilon_{i,t}) \tag{4}$$

$$TA_{i,t} = (\Delta CRA_{i,t} - \Delta CASH_{i,t}) - (\Delta CRL_{i,t} - \Delta SB_{i,t}) - \Delta DIS_{i,t} \tag{5}$$

其中，TA 表示总应计项；Δ REV 为两期营业收入的差额,：Δ REC 为两期应收账款的差额；PPE 是当期期末固定资产净值；A 是期末资产总额；Δ CRA 为当期流动资产与上期流动资产的差额；Δ CASH 为当期库存现金与上期的库存现金的差额；Δ CRL 为当期流动负债与上期流动负债的差额；Δ SB 是当期短期借款与上期短期借款的差额；DIS 是当期折旧数额。在计算 TA、Δ REC、Δ REV、PPE 都需要以上期期末资产进行标准化。随后，分行业对 TA 进行 OLS 回归，所得的残差为操控性应计利润 (DA)

## 6.2 描述性统计表 2

这是好公司和坏公司的描述性统计表。

表 7: 描述性统计表 2

| | lemon = 1 | | lemon=0 | |
|---|---|---|---|---|
| Variable | Obs | Mean | Obs | Mean |
| ch_sales | 5,044 | 3.755355 | 25,720 | -8.32032 |
| ch_assets | 4,949 | 1.352118 | 24,922 | 0.584309 |
| ch_ppe | 4,944 | 9.302766 | 24,904 | 63.73016 |
| ch_reg | 5,112 | 1.390063 | 26,764 | 1.416866 |
| wc_acc | 4,031 | -0.0031 | 21,154 | 0.001648 |
| ch_rev | 4,708 | 0.009332 | 23,904 | 0.012919 |
| soft_asset | 5,112 | 0.764053 | 26,764 | 0.754902 |
| da | 4,629 | -2.44E+08 | 23,402 | -3.26E+08 |
| ch_cs | 4,676 | 2.52E+09 | 23,645 | 6.61E+09 |
| ch_roa | 4,738 | 0.000656 | 23,951 | 0.006194 |
| lev | 5,112 | 0.56417 | 26,764 | 0.524005 |
| ch_emp | 4,723 | 1.130992 | 23,954 | 1.05698 |
| llemon | 4,742 | 0.540489 | 24,043 | 0.093666 |
| marketvalue | 5,112 | 8.22E+09 | 26,764 | 1.45E+10 |
| tobinq | 5,112 | 2.819399 | 26,764 | 4.33967 |
| booktomarket | 5,112 | 0.993255 | 26,764 | 1.051638 |

## 6.3 SMOTE 的伪代码

这是从 Chawla[5] 原文中抄录的算法：

Algorithm SMOTE(T, N, k)

Input: Number of minority class samples T; Amount of SMOTE N%; Number of nearest

```
neighbors k
Output: (N/100)* T synthetic minority class samples
1. (* If N is less than 100%, randomize the minority class samples as only a random
percent of them will be SMOTEd. *)
2. if N <100
3. then Randomize the T minority class samples
4. T = (N/100) * T
5. N = 100
6. endif
7. N = (int)(N/100)( * The amount of SMOTE is assumed to be in integral multiples of
100. *)
8. k = Number of nearest neighbors
9. numattrs = Number of attributes
10. Sample[ ][ ]: array for original minority class samples
11. newindex: keeps a count of number of synthetic samples generated, initialized to 0
12. Synthetic[ ][ ]: array for synthetic samples
(* Compute k nearest neighbors for each minority class sample only. *)
13. for i ← 1 to T
14. Compute k nearest neighbors for i, and save the indices in the nnarray
15. Populate(N, i, nnarray)
16. endfor
Populate(N, i, nnarray) (* Function to generate the synthetic samples. *)
17. while N != 0
18. Choose a random number between 1 and k, call it nn. This step chooses one of
the k nearest neighbors of i.
19. for attr ← 1 to numattrs
20. Compute: dif = Sample[nnarray[nn]][attr] − Sample[i][attr]
21. Compute: gap = random number between 0 and 1
22. Synthetic[newindex][attr] = Sample[i][attr] + gap * dif
23. endfor
24. newindex++
25. N = N − 1
26. endwhile
27. return (* End of Populate. *)
```

## 6.4 随机森林和 Bagging 算法的思想

bagging 的算法过程如下：

- 从原始样本集中使用 Bootstraping 方法随机抽取 n 个训练样本，共进行 k 轮抽取，得到 k 个训练集。(k 个训练集之间相互独立，元素可以有重复)
- 对于 k 个训练集，我们训练 k 个模型（这 k 个模型可以根据具体问题而定，比如决策树，knn 等）
- 对于分类问题：由投票表决产生分类结果；对于回归问题：由 k 个模型预测结果的均值作为最后预测结果。(所有模型的重要性相同)

随机森林根据下列算法而建造每棵树：

- 用 N 来表示训练用例（样本）的个数，M 表示特征数目。
- 输入特征数目 m，用于确定决策树上一个节点的决策结果；其中 m 应远小于 M。
- 从 N 个训练用例（样本）中以有放回抽样的方式，取样 N 次，形成一个训练集（即 bootstrap 取样），并用未抽到的用例（样本）作预测，评估其误差。

- 对于每一个节点，随机选择 m 个特征，决策树上每个节点的决定都是基于这些特征确定的。根据这 m 个特征，计算其最佳的分裂方式。
- 每棵树都会完整成长而不会剪枝（Pruning，这有可能在建完一棵正常树状分类器后会被采用）。

## 6.5 混淆矩阵

表 8: 混淆矩阵

| A Model | | prediction | |
|---|---|---|---|
| | | 0 | 1 |
| actual | 0 | TN | FP |
| | 1 | FN | TP |