财务诊断预测模型初步探讨

1. 建模目标

通过分析财务造假上市公司财务报告，预测给定上市公司是否财务造假。进一步，如果选取所有上市公司作为给定公司集合，筛选出其中更可能财务造假公司和更不可能造假的公司，通过回测比较两者之间是否有差异，评估模型效果。

1. 文献调研

通过查询财务安全或者财务预警模型相关资料，发现目前财务安全诊断主要使用单变量模型、多元线性判别分析、多元逻辑回归、多元概率比回归、人工神经网络等模型，而且研究目标大部分为判断上市公示是否会被特殊处理（ST）。根据调研结果，这些文章主要研究ST公司和非ST公司之间的财务报表中的区别，倾向于研究公司财务质量问题，和我们的研究对象不同。对于财务造假诊断，目前还没有相关研究。

1. 初步设想

首先收集被曝或被罚财务造假公司相关年度财务报告，从中提取相关指标。其次，将财务造假诊断问题看作一个分类问题，采用机器学习有监督学习的分类方法构建模型。最后，用模型预测所有上市公司的财务造假概率，并在其中划分为最不可能和最可能造假的公司两组，通过回测研究分类效果。

1. 模型构建及结果分析
2. 数据采集

首先在证监会及其下属证监局网站上找到受过行政处罚的公司名单，被处罚时间从2013年至今，一共64家公司。对于给定的公司名单，找到其造假年度财务报表，从其中提取相关财务指标，例如资产负债表（COM\_BALANCESHEETNAS）中的总资产、所有者权益，利润表（COM\_INCOMESTATEMENTNAS）中的营业总收入、营业利润、营业总成本，现金流量表（COM\_CASHFLOWSTATEMENTNAS）中的现金、存货的减少、预提费用的增加，主要财务指标表（COM\_MAINFINAINDEXNAS）中的净资产收益率、流动比率、速动比率等，总共116个，具体指标详见文件“指标选取.xlsx”。其次，对于有的公司造假年份不止一次的情况，根据不同的造假时间拆分出对应时间的年报数据作为一个样本。由于公司的数据库中最新的年报时间为2016年，因此选取造假时间为2015年及之前年份的财报数据作为训练样本。这样一共有120个样本，每一个样本对应一个公司和一个年度。

对于每一个样本，随机选择和它属于相同证监会二级行业的未造假公司作为对照样本，对照数据为这个公司相同年度的财报数据。由于每个公司的规模不同，公司的资产负债表、利润表和现金流量表中的数据均为绝对值，如果不做处理会对预测结果产生很大影响。因此将这三个表中的相关指标值除以该样本公司的总资产，用此值替换原样本中的对应指标值。最后，我们一共得到240个样本数据，包含120个造假公司和其对应的未造假公司的财务数据。由于财务数据中有很多的缺失值，因此我们选择在所有的样本中缺失值少于50%的指标作为最终的指标。

1. 模型建立
2. Logistic模型

首先采用这240个样本的年度财务指标构建logistic回归模型，同时使用L1范数来筛选指标。对于正则化参数的选择，我们采用交叉核实的方法，这个可以通过Python Sklearn的函数LogisticRegressionCV自动选择。为了评估模型效果，我们按照时间顺序对样本排序，将前80%的数据作为训练样本，剩下的20%作为测试样本，同时将AUC、Precision和Recall作为评估指标。为了防止随机选择的对照样本对预测结果的影响，我们分别进行10次试验，将10次试验的平均值作为最终的结果。结果如表1。

表 1 Logistic模型评估结果

|  |  |  |  |
| --- | --- | --- | --- |
| AUC | Precision | Recall | F1 |
| 0.5971 | 0.5243 | 0.5830 | 0.5583 |

其次，我们可以用Logistic回归模型预测所有2016年上市公司的财务造假概率。在公司的数据库中，一共有2512个上市公司的财务数据。这里我们用所有的240个样本构建Logistic回归模型，不再划分训练样本和测试样本。类似上面的分析，我们分别建立10个模型，每一个模型都随机选择对照样本，然后用这10个模型分别预测每个公司的造假概率，将所有模型结果的平均值作为该公司的财务造假概率。我们预测的所有公司造假概率分布直方图如图1所示。可以看出大部分公司都被预测为没有造假。对所有公司进行预测后，我们选择最可能是财务造假的200家公司作为一个组合（Most Negative），同时选择最不可能财务造假的200家公司也作为一个组合（Most Positive）。另外，我们从所有不可能财务造假的上市公司中随机选择200家作为另一个组合（Random Positive 200）。我们用聚宽对这三个组合进行回测，在2017年5月1日买入组合中所有股票，每只股票的买入额相等，然后在2018年4月30日卖出这个组合中的所有股票。回测结果如图1所示，基准指标为中证500指数。

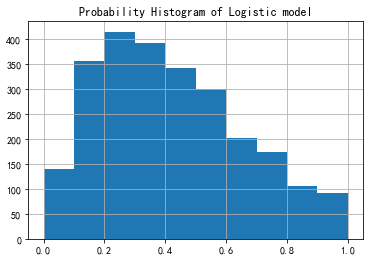


图 1 Logistic模型预测概率分布直方图

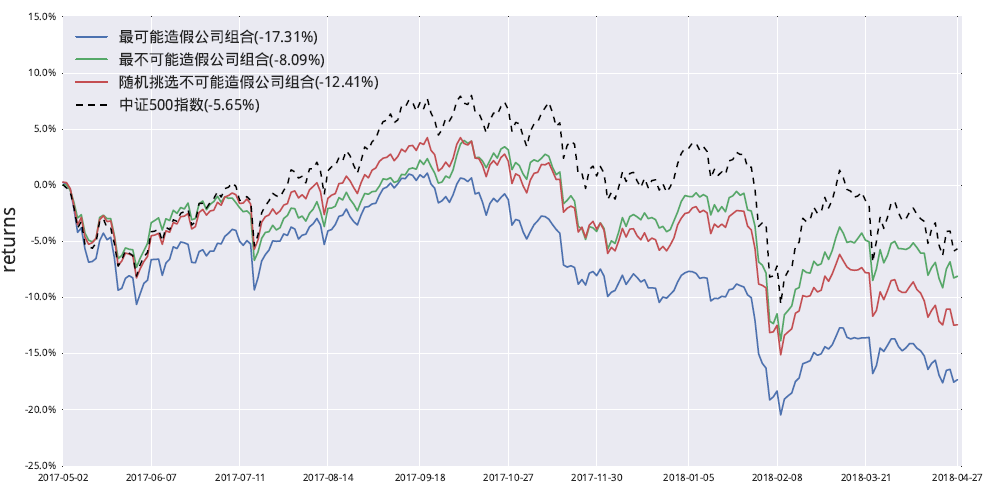


图 2 Logistic模型不同组合回测结果

1. 非平衡数据的Logistic模型

由于实际上市公司的数量远远超过被证监会处罚的公司数量，如果仅用财务造假的公司和相同数量的对照样本构造模型，可能会造成模型的欠拟合。因此，我们选择所有的上市公司2012年到2015年的财务指标数据以及所有的财务造假公司的财务指标作为样本，一共9257+120个样本。然后，我们同样采用Logistic回归模型对样本建模，但是与平衡数据Logistic模型不同的是，这里我们给予数据量小的造假样本更高的权重，给予数据量大的对照样本小的权重。权重的选择依赖于不同标签样本数量，而且权重之和为1。对于负样本，权重选择为9257/(9257+120)，对于对照样本，权重为120/(9257+120)。结果如表2所示。我们可以看到，相对于平衡数据模型，非平衡数据模型的各个指标都有明显下降，特别是精确度（Precision）和F1得分。这是由于非平衡数据模型预测中，过多的未造假样本被错误预测为造假样本。而AUC和召回率（Recall）在两个模型中并没有特别大的差别。

表 2 非平衡数据 Logistic模型评估结果

|  |  |  |  |
| --- | --- | --- | --- |
| AUC | Precision | Recall | F1 |
| 0.4917 | 0.0075 | 0.5243 | 0.0147 |

同样的，我们也可以用非平衡Logistic回归模型预测所有2016年上市公司的财务造假概率。我们预测的所有公司造假概率分布直方图如图3所示。大部分公司（73%）同样被预测为没有造假。对所有公司进行预测后，我们也按照预测概率将所有公司分为3组，进行了回测，回测结果如图4所示。可以发现，非平衡Logistic模型中的最可能造假组合和最不可能造假组合之间的收益差更大，说明了此模型能够显著区分不同财务类型的股票。

图 3 非平衡Logistic模型预测概率分布直方图

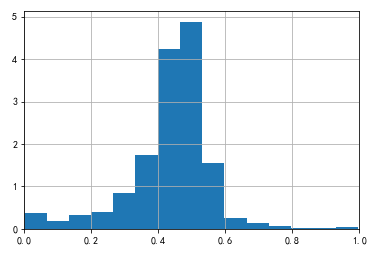
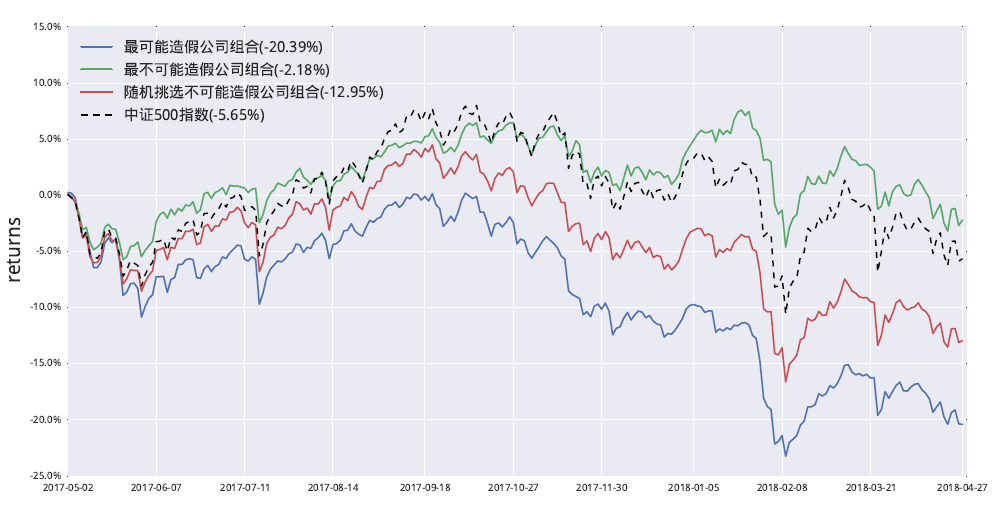


图 4 非平衡Logistic模型不同组合回测结果

1. LSTM模型

由于上市公司每年的财务数据和前几年的财务数据都有关系，因此我们探索使用造假公司前几年的财务数据对其财务造假进行预测。我们首先收集上市公司在造假年份前4年及当年的财务数据，然后运用LSTM模型对其造假概率进行预测。

由于LSTM模型使用梯度下降算法对模型进行优化，需要大量的数据，但是我们收集到的造假样本只有120个，如果直接使用这些样本建模会造成模型的过拟合。因此，我们需要对数据进行扩充。首先，我们可以按照前述方法随机构造对照样本。但是对每一个造假样本，我们随机选择10个对照样本，即将对照数据扩充10倍。对于每一个造假样本，我们对其每一个特征加一个正态分布噪声，均值为0，标准差为其实际标准差的0.1倍。对造假样本重复9次上述操作，加上原始样本，可以得到与对照样本同样数据量的负样本。通过这个方法，我们可以将原始数据扩充10倍。同样的，我们选择缺失值小于50%的指标作为最终指标，并且对资产负债表、利润表和现金流量表的数据进行处理。另外，我们对每一个指标的值标准化为0-1之间的值，防止指标的量纲不同对模型的影响。

在LSTM模型中，我们选择造假时间在2014年之前的数据作为训练集，2014年及其后的数据作为测试集。有些上市公司在造假年份并没有前4年的财务数据，因此我们需要在LSTM层之前加入Masking层，并将没有数据的那年的值都设置为一个Mask值，例如-100。当LSTM层发现输入均为Mask值时，会跳过这个输入。由于我们的数据量比较小，因此LSTM的单元数不宜过大，我们将其设置为16。为了防止过拟合，我们在LSTM和全连接层加入正则项l1\_l2，即ElasticNet正则化。同时，我们将训练数据分为训练集与验证集，当验证集的损失不在下降时，停止训练，防止由于模型过拟合导致模型泛化能力降低。整个LSTM模型参数如下：

Layer (type) Output Shape Param #

=========================================================

masking (Masking) (None, 5, 71) 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

lstm (LSTM) (None, 16) 5632

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

dense (Dense) (None, 1) 17

=========================================================

Total params: 5,649

Trainable params: 5,649

Non-trainable params: 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

注：第二行的71表示我们选择的特征数为71。

类似于Logistic模型，为了防止随机选择的对照样本对预测结果的影响，我们也进行10次试验，将10次试验的平均值作为最终的结果。结果如表2。同时我们也对2016年的上市公司财务造假情况进行了预测，概率预测直方图如图3所示。最后我们也按照预测概率将所有公司分为3组，进行了回测，回测结果如图4所示。

表 3 LSTM模型评估结果

|  |  |  |
| --- | --- | --- |
| AUC | Precision | Recall |
| 0.5939 | 0.6340 | 0.5280 |

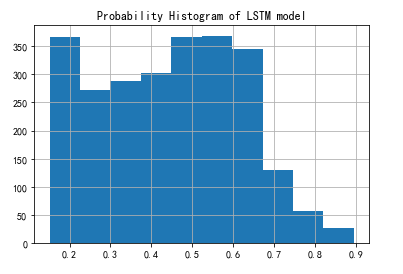
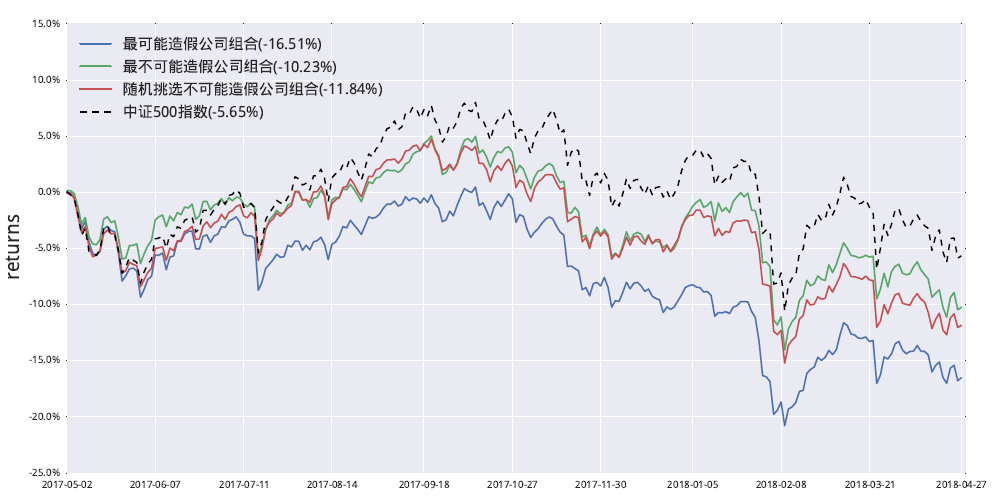


图 6 LSTM模型不同组合回测结果

图 5 LSTM模型预测概率分布直方图

1. 本福特定律

本福特定律是近来备受资本市场青睐的一种分析财务报告的方法。本福特定律因20世纪早期英国物理学家本福特而得名，其内容是：自然数据源（信用卡账单、采购记录、现金收据）生成的数字中，约有30%的数字的首位数是1，如1、1314；首位数为2的数字约有18%；顺序递减，首位数为9的数字少于5%。精确地数学表述为：在b进位制中，以数n起头的数出现的机率为。这个分布规律适用的数据集几乎无穷无尽，包括河流的长度、城市和国家的人口、证券交易所的成交量，当然会计数据（数据没有被人为操纵过）也同样适用。如果一组会计数据不符合本福特定律的话，就存在被篡改过的嫌疑。

首先收集所有财务造假公司财务报表中的所有财务指标，然后计算每个公司财报中的首位数字分布和本福特定律中的期望分布的欧氏距离，将这个距离看作为偏离本福特定律的程度。另外，同样分析未造假公司的财务报表中的所有财务指标，样本量为1200（即财务造假公司的数量的10倍）。不同类别样本偏离本福特定律的程度的分布如图7所示。可以发现，未造假公司的偏离本福特定律的程度更小。由于这两个样本均不符合正态分布，因此采用Wilcoxon秩和检验对这两个分组检验其是否来自于同一个分布，检验的P值为0.4692>0.05，不能拒绝原假设，即不能拒绝这两个来自于不同的分布。因此，在财务造假和未财务造假公司中，本福特分布并未有显著差别。

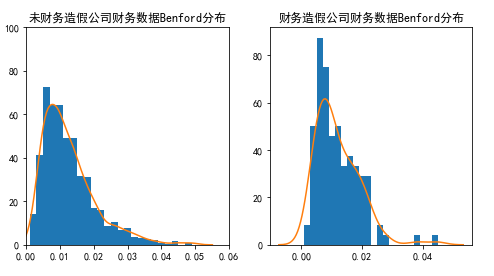


图 7 不同类别公司财务报表Benford分布

另外，可以采用KS检验方法，检验单个样本数据是否符合本福特定律。KS的计算公式为：KS=Max(|AD1−ED1|,|(AD1+AD2)−(ED1+ED2)|,…,|(AD1+AD2+…+AD9)−(ED1+ED2+…+ED9)|)，其中AD是统计数据的真实分布，ED为该数字在本福特定律中的期望概率。可以利用KS来检验该分布适合本福特定律，在5%的水平下，检验值为 1.36/srqt(N),其中N为该样本中的财务指标数量。如果该样本的KS值小于该检验值，可以认为符合本福特定律。对于造假公司，满足本福特定律的公司占所有公司的96.67%，而对于未造假公司，不满足本福特定律的公司占所有公司的93.67%。因此利用本福特定律判断上市公司财务造假情况可能会造成一些偏差。

接下来根据2016年财务报表偏离本福特定律的程度，将所有的上市公司排序，然后将偏离度最小的200家公司作为最不可能造假的公司，将偏离度最大的200家公司作为最可能造假的公司，回测其2017年度股票收益，结果如图8所示。可以看到，最可能造假的公司组合和最不可能造假的公司组合的回测收益都很差，而且最可能造假公司组合的收益要高于最不可能造假的公司组合。和前几种方法相比，本福特定律的效果明显差很多，说明此方法不一定适用于此类问题。

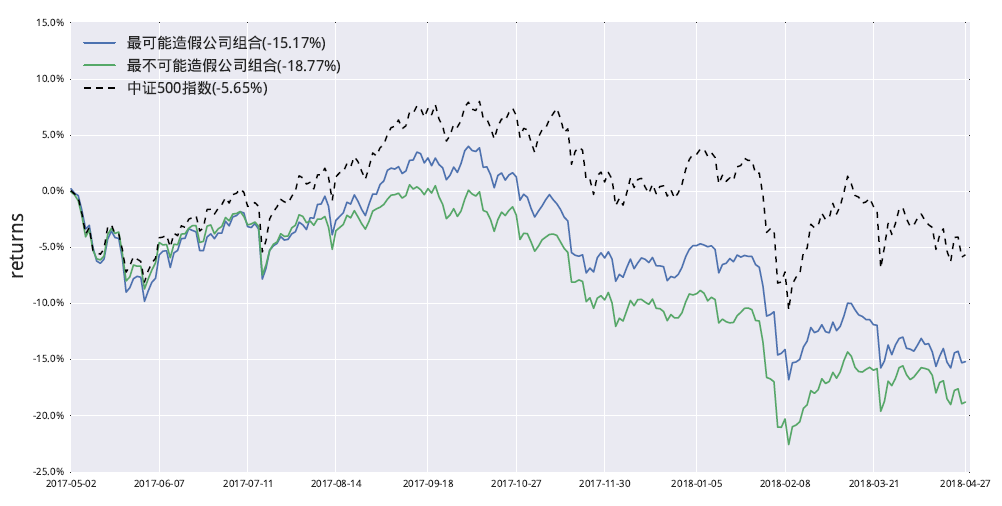


图 8 本福特定律模型不同组合回测结果