武汉大学《金融科技》课程作业 3

财务欺诈是一个全球性的重要问题。 如果不及时发现和预防，可能会对欺诈公司(eg. Enron)的利益相关者以及许多非欺诈性公司的利益相关者造成间接伤害。本次课程作业为：使用人工智能机器学习方法预测中国上市公司的财务欺诈行为，按照①数据收集—②数据存储和管理—③数据分析与处理—④数据可视化四步进行。

0. 作业要求

- 提交报告一份，格式自拟，保持简洁；提交代码（无需数据）。
- 采用坚果云链接提交作业，按提示输入学号姓名后上传文件即可，地址如下：

  https://workspace.jianguoyun.com/inbox/collect/b6b7177dd1864f70a1bff58f4fbba33e/submit
- 截止时间：4 月 27 日 12:15（根据最后提交版本的时间戳）
- 迟交规则：迟交一周扣当次作业的 25%

1. 数据收集

请按照以下路径下载预测需要的数据

1) 上市公司财务欺诈数据

从 CSMAR 数据库中下载财务违规信息，具体路径为：

公司研究系列 - 会计信息质量 - 财务报告信息 - 财务违规 - 上市公司财务违规表（日）

2) 上市公司财务数据

财务数据是公司状况的晴雨表，财务信息中包含了公司基本面的绝大部分情况。第二，基于公开可用财务数据的欺诈预测模型可以低成本应用于任何公开交易的公司。第三，现有会计文献中的大多数欺诈预测模型也依赖于公开可用的财务数据。

Bao et al. (2020)认为基于潜在的不完整行为理论将原始会计数据转换为有限数量的财务比率可能意味着有用的预测信息丢失。因此，在相同的方法下，原始财务数据和财务比率两种数据在预测中的性能也存在区别。（数据样本为全 A 股上市公司）

- 请按所示路径从 CSMAR 中收集以下 20 个原始财务数据。

利润表--营业收入/营业成本/营业税金及附加/营业利润/利润总额/净利润

资产负债表--交易性金融资产/应收账款/存货/固定资产/流动资产/总资产/流动负债/权益/应付账款/应交税费

现金流量表--现金及现金等价物/经营活动现金流/投资活动现金流/融资活动现金流

- 请按下列方法计算财务比率

ACC: 利润总额减去营业现金流，除以平均资产合计,平均资产合计等于 t-12 月份和 t 月份的总资产平均值

Change in receivables: 应收账款变动 / 平均资产合计

Change in inventory: 存货变动 / 平均资产合计

Soft Assets: (总资产-固定资产 – 现金及现金等价物)/总资产

Change in cash sales: 营业收入 - 应收账款变动

Percentage change in cash margin, where cash margin is measured as: 1 - (营业成本 - 存货变动 + 应付账款变动) / (营业收入 - 应收账款变动)]

Change in return on assets: (净利润 / 平均资产合计) - 上一期(净利润 / 平均资产合计)

Book-to-market: 权益 / 流通市值

2. 数据预处理

如果你已经下载并初步浏览了数据就会发现，数据并不像想象一样可以直接进行使用，因此请对你拥有的数据进行处理和简单的分析。

①数据清洗：请你整理出公司的财务违规年份和违规类型及相应处罚，生成一个新的数据集（每列对应：公司 ID、违规年份、违规类型、相应处罚），列名自取。

②请你合并欺诈数据和财务数据，分析财务违规的公司总体趋势，作图列出时序趋势，并展示数据的非平衡性。

③请你对两类公司（欺诈类/非欺诈类）收集的 28 个财务指标做描述性统计，要求包含'Mean', 'SD', 'Skew', 'Kurt', 'Min', '5%', '25%', 'Median', '75%', '95%', 'Max', 'n'。

④请你对比两类公司的区别，对两类公司的 28 个指标进行差分判断是否存在显著差异。

⑤请你选择一家出现过财务违规和一家没有违规的公司，对该公司的财务数据进行具体的分析。

3. 欺诈预测-Logestic 模型

延续 3，你已经做好了数据处理的前期准备，数据已经可以输入模型进行训练，那么哪个模型最适合欺诈预测是你需要探讨的问题。

一般情况下，对于因变量为 binary 信号(0 or 1)的计量问题，LR 模型有着天然的优势。

请你采用 Logestic 模型对上市公司财务欺诈进行识别，并对所有的自变量进行标准化处理。另外，采用拓展窗口进行模型训练，即使用最后 6 年的数据 2013 – 2018 作为测试集。例如,训练集是 1998 – 2011 测试集则为 2013 and 1998 – 2013 测试 2015, 训练集和测试集期间进行 2 年的间隔。

①请你对比两类数据（原始数据和财务比率）的 AUC 值。

②请你找到第 5 期模型的最优超参数。最优超参数选择——采用网格调参的方式进行最优参数的筛选，(hint:LR 参数主要调整优化算法，solver 只有五个可选参数，即 newton-cg, lbfgs, liblinear, sag, saga)

③logestic 模型的结果中，变量对于财务欺诈的预测如何起作用，请你找出 LR 模型变量对应的参数

4. 欺诈预测—RUSBOOST 非平衡算法预测财务欺诈

相信你在欺诈数据预处理的时候或许产生过这样的疑问，欺诈样本中信息太少了，面对数以千计的上市公司，财务欺诈和违规的公司数量非常缺乏，财务欺诈的样本和非财务欺诈的样本天然存在着极度不平衡，数据本身的非平衡性是否会影响预测的效果?

因此，除了使用 Logestic 模型外，请你使用 RUSBOOST 模型来解决欺诈样本的非平衡问题。RUSBoost 是 AdaBoost 的变体，它利用随机欠采样（RUS）解决类不平衡学习的问题。 它的工作方式与 AdaBoost 几乎相同，除了在每次迭代中执行 RUS 来解决欺诈性和非欺诈性公司的不平衡。

拓展窗口训练：采用 Logestic 模型中提到的扩展窗口进行训练。

最优超参数选择：采用网格调参的方式进行最优参数的筛选。首先在大规模的范围筛选 learning_rate [0.01, 0.1, 1, 10, 100] and n_estimators with [100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]. 在筛选出最优参数后请你再次进行进一步的筛选。例如,0.01 是最优参数，那么在同等数量级的范围内请再次进行网格调参筛选[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]，同样的 n_eatimators 也进行相同的处理 500 [400, 450, 500, 550, 600]

①请你找出 RUSBOOST 第 8 期学习的最优超参数。

②请你对比使用财务指标和财务数据得出预测的 AUC 值和平衡准确率，判断两者的学习效果，并绘制预测矩阵。

③请你对比 logestic 模型和 RUSBOOST 模型的预测效果，在给定公司的情况下，Logestic 模型和 Rusboost 模型的两种数据情形下预测的情况，能否识别出其欺诈（只需举一个公司的例子）。

④请你通过变量重要性找出对于财务欺诈预测最为关键某些的变量，并画热力图表示，并对比 LR 模型的预测参数。

参考文献：

[1] BAO Y, KE B, LI B, et al., 2020. Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach[J]. Journal of Accounting Research, 58(1): 199 – 235.