Data Analysis and Visualization

In class Lab: Case Study , Dec 15th, 2022

《Predicting Earnings Manipulation by Indian Firms Using Machine Learning Algorithms》

1. MCT 公司得到了更新数据，收集到了样本中全部公司的 ROE 的数据。请用线性回归模型，将数据集中的 ROE 作为响应变量 y，除 ID 和 Manipulator 之外的变量全部作为解释变量 X，进行线性回归，将回归模型的参数展示出来。

2. 基于上述模型，哪些变量对响应变量 y 的值影响比较大？模型的预测效果如何评判？当前模型拟合的效果是否良好？

3. 请用线性回归模型，将数据集中的 ROE 作为响应变量 y，除 ID 之外的变量全部作为解释变量 X，进行线性回归，将回归模型的参数展示出来。此时，Manipulator 这个解释变量，在模型中对响应变量 y 的估计效果显著吗？

4. 将除了 ID，Manipulator 和 ROE 之外的所有变量作为特征，将所有样本公司进行 K-means 聚类分析，聚成 2 个类。此时，每个个体被预测出的类别，与直接通过 Manipulator 这个变量的值进行对比，两者吻合程度如何？至此，你认为当前同为 Manipulator 的公司，会在其他数字特征上有一定的聚集效果吗？

5.将除了 ID，Manipulator 和 ROE 之外的所有变量作为特征，将所有样本公司进行层次聚类分析。你认为选取几个类别比较合适？不同的类之间分开的是否充分？此时，每个个体被预测出的类别，与直接通过 Manipulator 这个变量的值进行对比，两者吻合程度如何？至此，你认为当前同为 Manipulator 的公司，会在其他数字特征上有一定的聚集效果吗？

6. 至此，你认为只用当前数据，能够比较有把握的对那些操纵利润的公司进行预测和挖掘？如果不够，你认为还应需要知道哪些方面的数据信息？

7. 若给出了这些公司的年度财务数据，如季报，半年报，年报等等，你觉得还有什么模型可以帮助预测公司 ROE 的走势？