公司章程制度类公告占比相关因素的探究

1. 问题界定

与普通上市公司不同的是，新三板市场中的挂牌公司通常没有盈利要求，这对于在高速成长期的公司而言，不盈利的要求是具有合理性的。

除此以外新三板的公司在“章程制度建设”上，通常也是不够完备，这是因为挂牌的很多中小公司属于家庭式公司，其不太重视公司章程制度的建设。但是，从投资人的角度来看，一个公司的章程制度越完善，其对公司事务的处理越制度化合理化，越有利于公司的运营和发展。

因此“章程制度”越完善的公司才能更加获得投资人的青睐，章程制度的完善性成为了公司受投资人喜爱的一个重要因素。

本次项目的业务问题是：

探究与公司章程制度类公告占比相关的因素。

关于“章程制度”，新三板市场中的挂牌公司逐年披露的公告其中一类就是章程制度，我们可以将其占比作为目标变量，查看什么属性的公司与章程制度具有更强的相关性。

金融数据分析涉及到丰富的特征，通常是在建立模型时输入尽量多的指标，然后进一步利用模型对指标进行筛选，发现重要的相关因素。除去之前案例中使用的 49 个字段，在本次项目中我们引入一个新的数据表，表示不同公司的公告类别占比，见表“anno_cnt.xlsx”。

2. 数据准备

（1）导入 anno_cnt.xlsx 表后，请针对公司的章程制度类公告占比进行描述性统

计，并绘制直方图进行简单的说明。

（2）将“章程制度类占比”划分为占比高于占比低两类，作为模型的输出变量。

从业务的角度来看，投资人并不关注章程制度占比 3%和 5%公司之间的区别，因为占比上存在较小差距的公司其性质可能并不会相差很多，因此没有必要将实际的章程制度类占比作为最终模型的输出。

但是根据绘制的直方图，15%的公司与 5%的公司可能就属于完全不同的两种公司。因此，设定“章程制度占比”90%分位数作为阈值，将章程制度公告占比这一连续值离散化为一个分类变量，从而弱化类内的差距增强类间差距，更加明显地区分在占比上不同的公司。

（3）对模型的输入字段进行数据转换与数据分割，因为 y 变量的离散化是以 90%分位数作为阈值，导致正负样本比例为 1:9，请采用随机重采样实现数据重构。

3. 数据建模与结果解释

针对分类模型，我们已经学习了逻辑回归、KNN、朴素贝叶斯模型。其中逻辑回归是三种分类方法中唯一一种可以自动化稀疏变量的模型，本次项目中我们仍然使用逻辑回归模型。

（1）请参考中级案例一中的逻辑回归模型构建方法，根据模型的准确率与变量的稀疏度，利用网格搜索法确定最佳参数组合。

（2）在最优参数组合下，请绘制混淆矩阵或者 ROC 曲线评估模型结果，并根据非零特征及其参数对结果进行解释。

提高题：

在 python 的类 LogisticRegression 中，方法 fit 有另一个可调整的参数为

sample_weight，表示训练集中每一个样本的权重。权重越高的样本，在模型的评估结果中表现得越重要。一般来说，会根据正负样本的不同对 sample_weight 加以调整，使得分类模型更“看重”正类或负类样本。本次提高题中，请尝试设置负类样本权重默认为 1，修改正类样本权重在 1 上下改变，查看不同的 sample_weight 下，模型在测试集上精确率与召回率的改变。

提示：可以利用 sklearn 库中的类 confusion_matrix 计算精确率与召回率。