公司是否发生间接融资分类模型

1. 问题界定

在中级新三板的第一个案例中，我们已经大致了解了新三板这一公司股权交易平台的基本情况，包括中小企业获得资金的方式（间接融资、定向增发、股票交易），可能去向（上市、退市等），企业投资人面临的问题（公司质量堪忧、投资盈利模式不清）等等。

针对投资人而言，亟须理清与公司发展趋势、潜在价值相关的因素，从而为后续的投资策略提供参考。在课程案例中，我们通过建立逻辑回归模型，分类公司是否会发生股票交易，作为一种公司价值的体现，而在本次项目中，我们从另外一种公司获得资金的角度出发，分析公司的潜在价值：

根据公司特征分类公司是否会发生间接融资。

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。总体数据的基本情况与提取的 49 个字段已经在课程案例中介绍过，在项目中依然沿用这些数据。

2. 数据准备

（1）除去字段 company_code，请对其余 48 个字段做基本的描述性统计，如最大值、最小值、分位数等（因为字段数较多，直接在 jupyter notebook 中输出可能不太好查看，可以将统计结果输出到 excel 表中）。针对与是否发生融资相关的字段进行说明。

（2）绘制现有字段的直方图，并判断字段类型和直方图形状判断其大致属于什么数据分布。

（3）除了“数据字段处理记录.xlsx”中标灰的 7 个字段，其余 41 个字段中存在分类变量取值过多的情况，根据连续型/离散型确定缺失数据的填充方法、异常值处理方法等对

这些字段进行处理。

注：可以参考课程案例中的思路，规定某些字段的取值正常范围，然后处理字段。

提高题：

在新三板案例中，数据异常值的处理主要是人为划定字段的正常取值范围，是一个需要一定业务经验且相对主观的判断。可以根据字段的极值情况、分位数值尝试合理地修改现有的异常值处理方法，观察对后续的建模效果是否有影响。

请根据业务问题对字段进行数据转换。

注：①对现有字段转换得到模型的输出变量，即是否发生间接融资；②文本型字段的转换。

（5）KNN 算法构建分类模型

本次项目中的业务问题和课程案例中相同，也是一个分类问题，在此我们使用 KNN 算法构建分类模型（下文会对算法原理进行介绍）。和逻辑回归模型相似，KNN 算法要求特征符合标准化后的近似正态分布。请根据模型特点，确定需要进行怎样的数据转换、数据重构以及数据分割。

3. 数据建模与结果解释

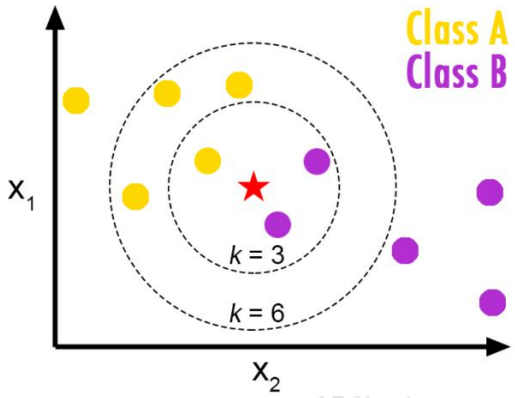
KNN（K-Nearest Neighbor，也称 k 近邻）算法是一种常用的监督学习方法，其工作机制可以简单概括为“近朱者赤，近墨者黑”。即给定测试样本，基于某种距离度量找出训练集中与其最靠近的 k 个训练样本，然后基于这 k 个“邻居”的信息来进行预测。通常，在分类任务中，可以选择这 k 个样本中出现最多的类别标记作为预测结果（也称为“投票法”）。

KNN 算法是一种非参数模型，也就是说模型训练结果并不具有参数，训练阶段仅仅是把样本保存起来，测试时，输入实例，将新实例的每个特征与样本集中的数据对应的特征进

行比较，计算距离，找到与该实例最邻近的 k 个样本，这 k 个样本的多数属于某个类，则这个输入实例就属于某个类。

影响 KNN 算法效果的主要参数有 k 值，以及距离计算方法。k 值决定了类别判断的依据样本数，如下图所示。

当取 k=3 时，星型实例被判定为 B 类；当 k=6 时，实例被判定为 A 类。



常用的距离计算方法：欧式距离、曼哈顿距离。

欧式距离：样本之间的距离 $D = \sum_{i=1}^{p} (x_i - y_i)^2$

其中 X 表示输入实例，Y 表示训练集中的样本，p 表示样本的特征数（欧式距离也被称为 l2 距离）。

曼哈顿距离：样本之间的距离 $D = \sum_{i=1}^{p} |x_i - y_i|$，也被称为 l1 距离。

基于距离的计算方法，KNN 算法要求特征符合标准化后的近似正态分布（在不严格的情况下部分分类变量可以纳入计算，但是会影响结果）。

通过 sklearn 库中的类 KNeighborsClassifier 实现 KNN 算法，k 值和距离计算方法分别通过参数 n_neighbors, p 来设置。其中 p=1 表示采用 l1 距离，p=2 表示采用 l2 距离。分类算法效果的评估可以采用准确率和混淆矩阵的方法。

（1）请根据 KNN 算法的介绍，构建分类模型，并尝试不同的参数组合，通过准确率

确定最优参数组合。

（2）请根据确定的最优参数组合，绘制模型基于测试集的混淆矩阵和 ROC 曲线，并对结果进行解释。

提高题：

计算最终 KNN 模型的精确率和召回率，并结合业务对两个指标进行解释。