公司发生交易天数预测

1. 问题界定

在中级新三板的第二个案例中，我们已经通过建立线性回归模型，预测公司的间接融资金额，从融资方面评估公司的价值水平。

而在本次项目中，我们从股票交易的角度对公司价值进行预测，具体的业务问题是：根据公司特征预测公司发生交易的天数。

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。

本项目中依然沿用这些数据。

2. 数据准备

（1）除去字段 company_code，在对其余字段进行基本的描述性统计后，请针对与交易天数相关的字段进行说明。

（2）选取交易天数不为 0 的数据作为后续分析的内容，并绘制现有字段的直方图。请比较当前数据子集的直方图与原数据集直方图，判断字段的数据分布是否有明显的变化。

（3）在本次项目中，仍然选择线性回归模型，因此输入数据需要符合正态分布。请参考之前案例，确定数据字段处理记录.xlsx 中字段的预处理方式。

（4）在数据建模之前，通过 train_test_split 划分训练集和测试集。

3. 数据建模与结果解释

线性回归在 python 的工具包中有多种实现的方法。除了在案例二中介绍的 sklearn 库，statsmodels 统计分析包中也有实现线性回归模型的类 OLS。与 Lasso 不同，类 OLS 实现的是优化函数中不包含正则化项的，最简单的一类线性回归模型。类 OLS 没有需要调整的

超参数，只需要通过类中的 fit 方法拟合模型，results.summary 查看拟合结果即可。

可参考 statsmodels 官方文档：http://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.html#statsmodels.regression.linear_model.OLS

或博客介绍：https://blog.csdn.net/qq_17119267/article/details/79108333

（1）请根据类 OLS 的使用教程，先将所有的特征输入模型中，构建线性回归模型，并通过 t 检验得到对因变量有明显影响的自变量。

t 检验判断的方法：

查看 results.summary 中列 P>|t|的取值，小于 0.05 表明对因变量有明显影响。

（2）保留有明显影响的特征，再次进行拟合，查看 P>|t|值，并对结果进行解释。

注：经过 onehot 编码后的变量会变成多个变量，新变量个数与原变量取值数相同，在保留特征时注意考虑这部分的影响。即变量 transfer_mode 因为有 2 种取值，在拟合结果中代表 x1-2，其他类似变量同理。

提高题：

在以交易天数为因变量进行模型拟合时，发现 listing_days 和交易天数有较强的正相关性，但是这种正相关性不具有有效的含义。请针对这种问题，采取新的方法构建模型，使得模型可以更好地评估公司在股票交易方面的价值。