题目 1. 关于证券交易的时间序列聚类和采样频率分析

问题背景

许多古典的投机策略习惯观察 3-5 日的日频 K 线，预测次日涨或跌。同时，也有许多交易策略关注长期收益，以日、周、月为单位买卖。每个股票在不同频率尺度下，能使用同种预测模型吗？在不同的时期，同种股票的日频交易，可以用同一个模型预测吗？请对此展开研究。

数据集和资料

数据集 `https://finance.yahoo.com` (获取方法参考 Python 函数库 `yfinance`)

请注意时间序列预测模型和时间序列的外部回归之间的区别。时间序列预测模型是指通过一个时间序列的以往数据预测未来数据，而时间序列的外部回归是指拟合一个时间序列本身和一个分数，这个分数并不来自时间序列中的一个值。在研究过程中，除课程内容以外，建议了解华为香港诺亚实验室发表的短时间序列预测模型 BHT-ARIMA 和时间序列特征自动提取函数库 TSFRESH，这些内容可能对你完成本项研究有所帮助。

中频数据集 1：请下载中证 50 指数成分股在 2021 年**其中 15 个交易日**的日频收盘价数据。确定好日期后，所有证券都应该来自这一天，并且去除停牌的股票。

中频数据集 2：请在中频数据集 1 的基础上，下载几个月以后某 15 个交易日的日频收盘价数据。

低频数据集：请下载这些股票在 2021-2022 年所有的周频、月频收盘价，若证券交易所有收盘集合竞价，不需要特别去除，请按默认设置获取数据。如果股票在这一年发生过除权，请使用前复权数据。

研究任务

问题 1-1. 在 3 个数据集上计算对数收益率，建立一个时间序列预测模型，预测对数收益率。不需要使用交叉验证，但要分为 70% 的训练集，10% 的验证集，20% 的测试集。

问题 1-2. 请分别优化 3 个时间序列预测模型，搜索模型的超参数，以及自变量的项数（移动窗口长度）。比较上述 3 个模型的结果，评价采样频率对时间序列预测的影响。由于时间尺度不同，直接比较 MAE 是不科学的，请思考合适的模型评价指标。

问题 1-3. 用中频数据集 1 训练好的模型，在中频数据集 2 上预测，与问题 1-2 中频数据集 2 上的预测结果有何不同？这对量化金融模型的建模师有怎样的启示？