高频交易 (High Frequency Trading) 是一种新型金融交易模式，使用高频率的算法交易进行实时分析和快速交易。

具有以下一些特点：

交易频率高：多策略多品种，每天交易次数多达上万次

持仓周期短：日内交易，不超过一天

盈利稳定：实际交易中的最大回撤率低

收益率高：资金量少的（千万级）年化收益可高达300%

总体来说，高频交易比传统交易更快、更准确，使得交易商更容易获取差价收益，减少交易中出现错误的几率。量化分析模型在高频交易中被广泛应用，提高高频交易的精度和操作效率。

目前主要应用范围：股票交易，期货、外汇、债券等金融市场

【国内外研究】

量化交易在国外已经发展40多年，已趋于成熟，交易品种相对多样交易市场完善。机器学习算法在国外量化市场上的应用已经很广泛。国外相关研究主要集中在机器学习在股票市场中的应用，少部分聚焦于期货市场研究。研究早期的算法主要以鲁棒性较好的支持向量机、随机森林、梯度提升树等算法为主。近些年来，伴随着深度学习的发展，有关LSTM算法的研究越来越多，但也基本集中于股票市场。

纵观国内外的文献研究可以发现，金融数据序列预测数据早年的研究算法以线性预测和传统的机器学习为主。传统的线性预测方法，如自回归模型（AR）、滑动平均模型（MA）、自回归滑动平均模型（ARMA）都是基于历史序列数据、白噪声或两者组合的线性预测方法，但要求时间序列是平稳的或差分序列是平稳的，因此存在一定的局限性。

相较而言，机器学习算法，如支持向量机、随机森林等，不要求序列的平稳性，可以接受序列本身的统计特征，并将外部环境特征引入模型；然而，处理外部环境对序列的影响需要人工构造特征，进行特征筛选，效率相对较低。近年来，模型的选用主要集中在神经网络及其优化。不需要人工构造合适的输入特征，在隐藏层中实现了有效量化因子的寻找，模型预测效果通常优于传统机器学习模型的预测效果。

【研究路线】

我们项目的研究可分为理论研究和实盘检验两个部分。

1）在项目中期，主要探索高频量化的各类技术方法；了解了传统的市场因子构建以及各类经典模型。

2）项目中后期，我们根据企业方提供的数据集进行实盘验证，主要对比的是经典的深度学习模型LSTM和时间序列模型ARIMA；两类模型都是常用的量化交易模型，各有优势。主要利用模型在数据集上预测具体的价格和涨跌方向，进行对比分析。

数据集用的是：

【研究创新点】

项目的整体创新主要分为三点：

1）数据集：过往的研究集中在股票、期货；而我们应用的数据集是冰宽量化提供的数字货币交易数据集，存在一定场景差异。

2）模型对比：对比了两类经典模型（LSTM & ARIMA）并进行评估分析

LSTM（深度学习模型）：引入市场因子

ARIMA（时间序列模型）：引入动量因子、反转因子

在一般模型的基础上，创新性引入因子策略进行研究

预测两方面的内容：具体数值 & 涨跌方向

实际意义：

1）探讨高频交易技术在量化对冲基金中的应用，分析其优缺点及其适用的范围。

2）理解量化对冲基金的运作机制，为量化对冲基金投资者提供更多的信息，帮助其更好地理解和把握市场风险。

3）研究结果对于投资机构和金融机构具有参考价值，可以帮助其制定更有效的投资策略和风险管理

### LSTM文稿

1. 下面为LSTM模型与实现，分为背景介绍，LSTM模型介绍，市场因子介绍和模型实现及分析四个部分
2. 高频交易的传统处理方式中，线性模型是常用的方式之一；然而该模型较为简单，难以捕捉不同因子与预测目标间较为复杂的非线性关系；而可以处理非线性关系的时间序列模型中，循环神经网络 RNN 如下图所示，其中每一时刻的值取决于上时刻结果和本时刻输出值。但RNN容易出现长期依赖问题，即在做反向传播时序列早期的数据难以被使用，因此难以捕捉高频交易中市场的长期趋势。因此使用传统方法和RNN不能满足需求
3. 而作为 RNN 的改进版长短时间记忆模型 LSTM 可以有效解决这个问题，其结构如图所示。其中复杂的多门结构和非线性激励函数使得该模型可以表征高频交易市场中相关因子和计算目标间复杂的非线性关系；同时，对于高频交易市场而言，单元状态更新过程中的对应元素相加，使得可以通过调整超参数缓解长期依赖问题，因此可以更好的学习到市场的长期趋势
4. 作为输入的市场因子，通过查阅文献，我们提取出三个市场因子。VOI 因子计算式如右图所示，即当买卖价不等时，需要计算双方交易量的差值；该因子可以反应交易量的不平衡，其可以同时捕捉交易量和交易方向两个维度，从而更清楚展现了市场的总体情绪和走向；而对于OIR因子，其计算式如下所示，通过计算买卖价格在双方市场中的占比可以考虑到买卖委托量本身规模的大小，从而衡量订单不平衡；而 Bid-Ask Spread 因子衡量了市场的流动性。上述因子在后续中会作为输入；
5. 在模型实现中，为了更好的符合高频交易市场特性，我们将任务分为具体交易值预测和涨跌幅预测，同时通过超参数寻优后，我们将神经网络设为3层，其神经元数量为 [50, 10, 1]，学习率为10^{-4}，批量大小为40，同时使用交叉验证，验证集占比为0.2。模型结果表现如右所示，其中具体交易值预测任务中与具体值有一定差别，而涨跌符号预测任务中测试集较高，准确率达到0.831，即涨跌符号预测任务表现更好。此可能与高频交易市场中噪声较大，而预测涨跌符号的二分类任务对噪声更不敏感。

ARIMA文稿

* ARIMA 模型是一种基于时间序列分析的预测模型，它通过对时间序列数据进行分析和建模，预测未来的值。ARIMA 模型可以看作是 AR（自回归）模型和 MA（移动平均）模型的结合体，它将这两种模型中的优点结合在一起，同时消除它们各自的缺点。ARIMA 模型可以用于处理非平稳时间序列，其中包括趋势、季节性和周期性 等多种非平稳特征。ARIMA 模型可以通过对时间序列数据进行差分操作，将非平稳序列转化为平稳序列，然后建立模型进行预测。
* 模型步骤：数据预处理：对时间序列数据进行预处理，如去除异常值、缺失值等。 平稳性检验：使用 ADF 检验、KPSS 检验等方法检验时间序列数据是否为平稳 时间序列，若不是，则进行差分操作。自相关和偏自相关函数图（ACF 和 PACF）的分析：根据自相关和偏自相关函数图以及AIC和BIC来确定自回归阶数 𝑝 和移动平均阶数 𝑞。
* 接下来介绍下结果：本研究以研究过去时刻的收益与随机波动因素对于之后收益的影响。取前 1000 个数据作为研究对象，在 t 时刻，通过对于 𝑡 − 𝑘1 到 𝑡 时刻的收益作为动量因子、𝑡 − 𝑘2 到 𝑡 时刻的随机波动作为反转因子，用 ARIMA 模型预测 t 到 t + 1时刻的收益..
* 首先绘制了右边的ACF与PACF图像，由右图可发现 ACF 在 1，11 位置处显著，PACF 在 1，2，3，4，5，6，8，9，10， 12，31，36，38 位置处显著。因此选定 p 和 q 的范围分别为 1-38，1-11，遍历 38\*11 种模型，选择 BIC 最小者作为结果。
* 最后构建 ARIMA（1，0，1）模型，结果如黑色图所示
* 做出预测值与真实值的对比图像、残差图（图 12b）以及残差概率密度分布（图 12c）。由上图结果可以得出如下结论：预测值与真实值的折线图具有一定的重合度，计算两者的相关系数为0.81506057， 为强正相关，模型趋势拟合较好； • 除了初始的一段时间，之后的残差均匀分布在条形宽带中（由于模型基于过去数据预测，对于初始数据的预测缺少相应的动量因子和反转因子）；残差的概率密度分布近似为均值为 0 的正态分布。利用此模型进行预测，观察 该模型十分合理。 最后计算预测值与真实值的均方误差、均方根误差、平均绝对误差、𝑅2 等定量 指标对预测模型进行评估。结果为均方误差: 0.00000001，均方根误差: 0.00011396， 平均绝对误差: 0.00007994，𝑅𝑠𝑞𝑢𝑎𝑟𝑒: 0.30498027 发现预测结果较为良好。

补充：时间序列动量效应指证券若在历史一段时间内获得较好（差）收益则未来其仍能继续 获得显著高（低）收益的现象。时间序列反转效应是指证券若在历史一段时间内获得较好（差） 收益则未来其收益会显著变差（好）。投资者可以依据这一规律对股票的未来收益进行预测，构建动量和反转策略。构建策略的关键之一是探索如何选取历史时间窗口的长短。尝试构建日内（即因子的构建和预测所用的数据均为同一天）动量/反转因子，在t时刻用t − k到t时段的数据预测当日t到t + 15秒的收益。探索时间窗口的长短即参数k的取值对预测效果的影响。