商品期货涨跌幅预测问题

1 摘要

随着金融市场的全球化和信息技术的快速发展，商品期货价格涨跌幅预测对投资决策和风险管理具有重要意义。本文基于 2017 至 2025 多年商品期货的主力合约等数据，构建了基于长时间序列预测模型 LSTM 深度学习预测框架，为商品期货市场的投资决策提供支持

对于数据预处理和特征提取，本文首先对各主力合约的收盘价，交易量进行可视化分析，为特征提取提供数据支持。随后，剔除诸如 exchange 等关于涨跌幅预测不相关的数据。接着，利用四分位数检测数据 $close$，$volume$ 数据中的异常值，并采用线性插值处理。最后基于原始数据，通过相关性分析，提取与涨跌幅相关性强的以下 3 个特征：交易量随时间的变化率（$volume\_change\_rate$）:交易量是市场活跃程度的重要指标，依据道氏理论和量价分析理论，价格变动若伴随交易量显著增长，则趋势更可能持续，交易量突增常伴随着主力资金的进出或情绪突变; 交易量滞后特征（30min&1d）（$volume\_lag$）：滞后交易量可以反映市场在过去某一时刻的活跃程度，有助于捕捉市场的短期记忆效应与行为惯性。30 分钟滞后反映了短周期的交易节奏，1 天滞后则体现了日内波动对次日走势的影响; 交叉特征（波动率 × 交易量）（$vol\_crossfeat\_volume$）：波动率衡量市场的不确定性，交易量反映市场的活跃度。两者的交叉特征可揭示市场剧烈波动前的征兆。

对于模型选择，题目是典型的监督学习 + 长时间时间序列回归问题，有强烈的时间依赖性以及非线性特征影响。而基于 RNN，LSTM 有如下核心优势：解决长期依赖问题：LSTM 通过门控机制和记忆单元结构，能有效捕捉数百个时间步长的依赖关系，克服了传统 RNN 的梯度消失问题; 选择性记忆与遗忘：三种门控机制（遗忘门、输入门、输出门）使 LSTM 能智能地过滤信息，同时保留长期趋势和响应短期变化; 多尺度模式同时捕捉：能够在同一模型中捕捉长期趋势、中期周期和短期波动，特别适合金融等存在多时间尺度特征的应用场景。

对于模型实现，首先输入数据通过 3 个模块：MinMaxScaler（归一化特征缩放器）将特征缩放至 [0,1] 区间，消除不同量纲的影响；Cleaner（数据清洗器）处理缺失值和异常值；TS Dataset(时序数据集构建器) 通过滑动窗口法生成具有固定时间步长（30 个时间点）的序列样本。网络架构由三个主要模块组成：Inputer(输入处理模块)，包含 BatchNorm1d 层，提高训练稳定性；LSTM 核心架构，采用双层设计（64-32 隐藏单元），形成深层特征表示；Outputer(输出处理模块)，包含全连接层和激活函数，将 LSTM 提取的特征映射为未来 30 分钟的涨跌幅预测值。

对于模型实现，首先输入数据通过 3 个模块：$MinMaxScaler$（归一化特征缩放器）将特征缩放至 $[0, 1]$ 区间，消除不同量纲的影响；$Cleaner$（数据清洗器）处理缺失

值和异常值；$TSDataset$（时序数据集构建器）通过滑动窗口法生成具有固定时间步长（30 个时间点）的序列样本。网络架构由三个主要模块组成：$Inputer$（输入处理模块），包含 $BatchNorm1d$ 层，提高训练稳定性；LSTM 核心架构，采用双层设计（64-32 隐藏单元），形成深层特征表示；$Outputer$（输出处理模块），包含全连接层和激活函数，将 LSTM 提取的特征映射为未来 30 分钟的涨跌幅预测值。

对于模型的预测效果，本文通过多种可视化方法展示了预测效果。首先，绘制了预测收盘价与实际收盘价的时间序列对比图，直观展示模型在不同市场阶段的预测准确性；接着计算预测的涨跌幅并做出可视化。最后，采用方向预测准确率（Direction Prediction Accuracy）及涨跌幅预测 MAE 对预测效果进行评估

2 问题重述

2.1 问题背景

商品期货（如螺纹钢、铁矿石、焦炭、焦煤等）是金融市场中的重要交易品种，其价格波动受到多种因素的影响，包括供需关系、宏观经济政策、国际市场变化等。若能利用历史数据预测商品期货未来的涨跌幅，则可帮助投资者更好地进行交易决策。

2.2 问题提出

现有数据集为 1 分钟级数据，包括时间戳、开盘价、最高价、最低价、收盘价、成交量、持仓量等。请基于该数据集建立数学模型，预测商品期货未来 30 分钟的涨跌幅。涨跌幅定义为涨跌幅 $= \frac{Pt+30-Pt}{Pt} * 100\%$ 其中 P_t 是当前时刻的价格，P_{t+30} 是 30 分钟后的价格。要求从 1 分钟级数据中提取出可能影响 30 分钟涨跌幅的特征，选择合适的机器学习模型对未来 30 分钟的涨跌幅进行预测。解释模型的选择理由，并使用适当的评价指标评估模型的性能，讨论模型的局限性及可能的改进方向。

3 模型假设

在建立预测商品期货未来 30 分钟涨跌幅的数学模型前，需对问题作出合理的建模假设。本文作出如下模型假设：

1. **市场具有短期可预测性：**
   假设商品期货价格在短期（如 30 分钟）内的波动具有一定规律性，可以通过历史的价格、成交量、持仓量等数据进行建模与预测。虽然市场整体是弱有效的，但在微观时间尺度上存在短期模式或信号。

2. **历史数据中蕴含未来信息：**
   假设过去一段时间内的交易数据（如过去 30 分钟的价格和成交行为）中包含了对未来价格变动趋势的有效信息，机器学习模型可以从中提取出这种映射关系。

3. **数据是按时间顺序生成且无信息泄漏：**
   假设训练、验证和测试数据均按时间顺序划分，未来数据不会出现在训练样本中，确保模型不利用“未来信息”来预测。

4. **价格波动主要受内部因素驱动：**
   初步假设模型只考虑交易数据本身（如价格、成交量、持仓量等），未纳入外部宏观因素。即，短期内商品价格波动主要由市场自身行为决定。

5. **特征变量之间相互独立或弱相关（用于部分模型）：**
   对于一些机器学习模型（如线性回归、决策树等），默认特征之间不是高度共线的。若存在强相关性，应通过降维或正则化处理。

6. **无重大政策或突发事件扰动：**
   假设模型训练和预测的数据段未处于特殊时点，如重大政策发布、突发灾难、战争等极端事件导致市场失真，这种情形应排除或特殊建模。

7. **数据采集频率与市场反应一致：**
   假设 1 分钟级别的数据能够捕捉市场行为的主要变动特征，且不会错过关键的市场信号，适用于建模 30 分钟后的涨跌幅。

8. **标签构造方式合理且滞后窗口固定：**
   假设涨跌幅的定义方式为

$$\text{涨跌幅} = \frac{P_{t+30} - P_t}{P_t} \times 100\%$$

   是一种有效衡量未来价格变动的方法，并且“30 分钟”是一个合理的滞后窗口长度，符合常见交易策略的时间尺度。

4 问题求解

4.1 数据预处理

预处理 preprocess 的核心: 将数据从以时间为分类标准变为以期货类型为分类标准

1. 去掉和文件名时间不相同的所有数据, 保证仅包含当天的数据

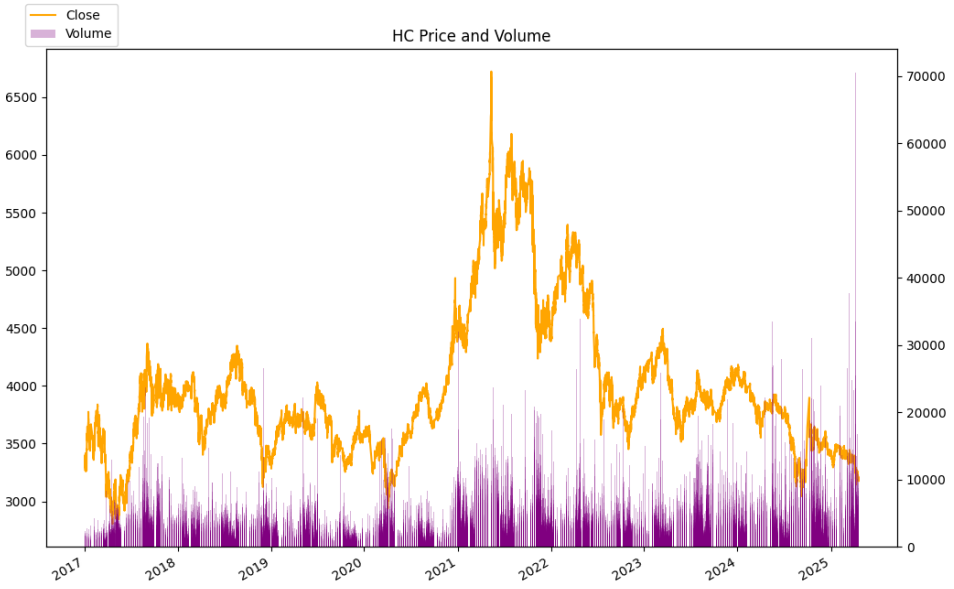
2. 去掉 exchange,contract,symbol,open,high,low,openinterset 这些与涨跌幅不相关的数据

3. 检查 close 是否是 float64 类型,volume 是否是 int64 类型, 如果是字符串类型则需要进行修改

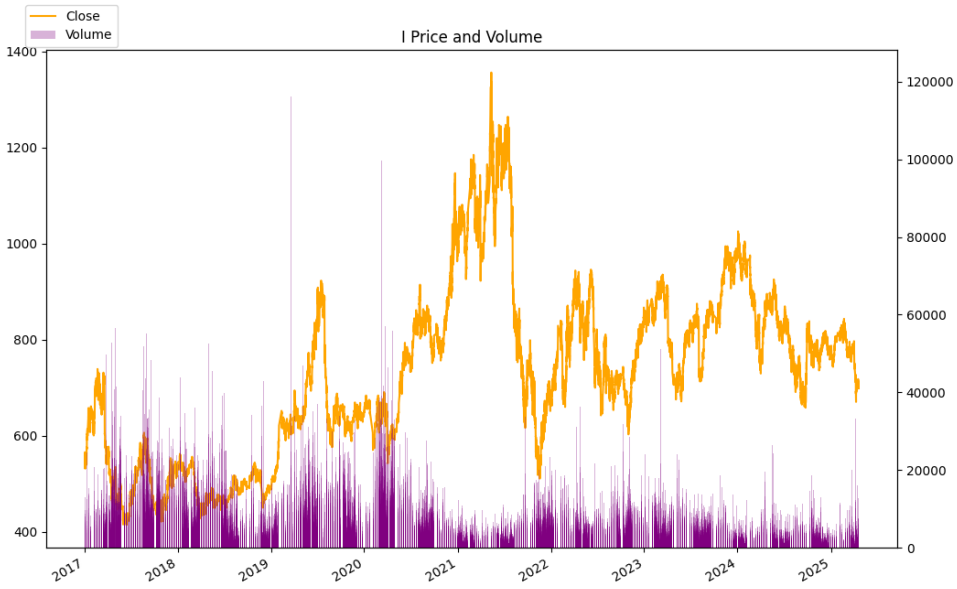
4. 四分位数法检查 close 和 volume 数据中的异常值, 出现异常采用线性插值法进行平滑处理

4.1.1 最终得到仅包含 datetime-close-volume 的 7 个数据文件: 此处篇幅原因暂时仅给出 3 张。

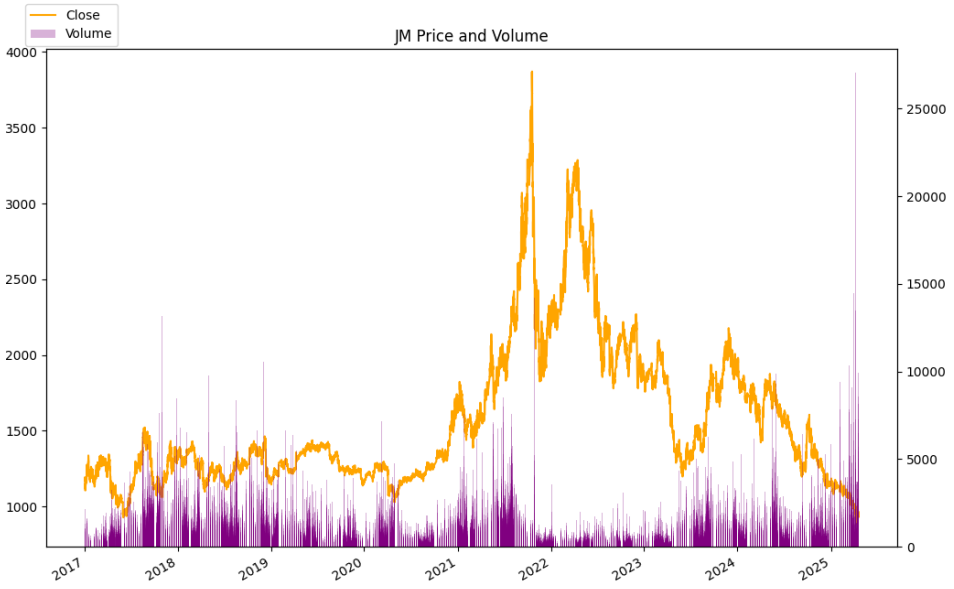
1. 给出异常值处理前的 volume 和 close 的重叠折线图



异常值处理前 HC 的 volume 和 close 的重叠折线图



异常值处理前 I 的 volume 和 close 的重叠折线图



异常值处理前 JM 的 volume 和 close 的重叠折线图

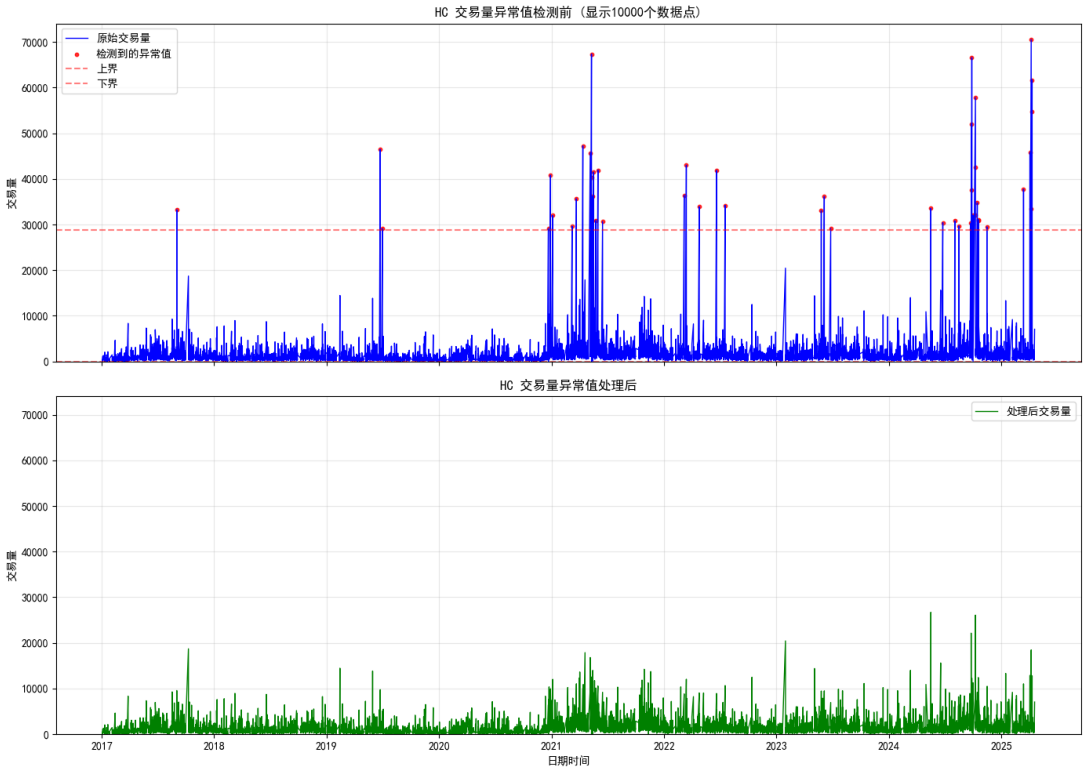
2. 给出异常值处理后的 close 随时间变化的数值折线图:

异常值处理后 HC 的 close 的折线图

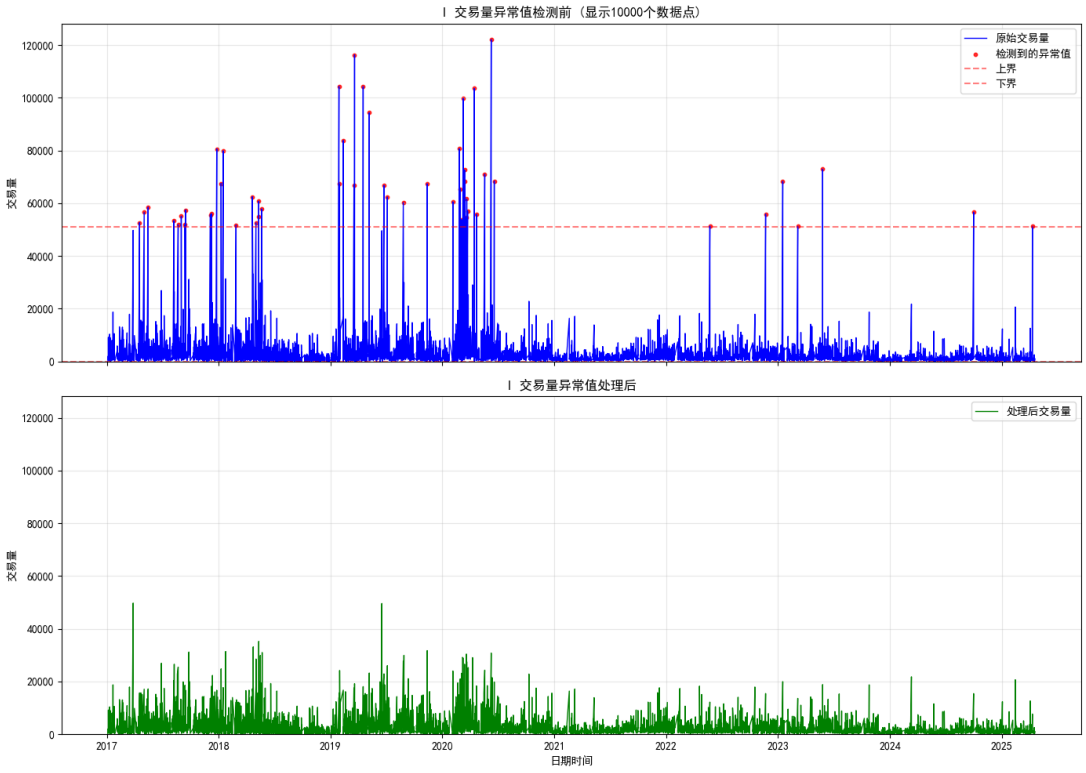
异常值处理后 I 的 close 的折线图

异常值处理后 JM 的 close 的折线图

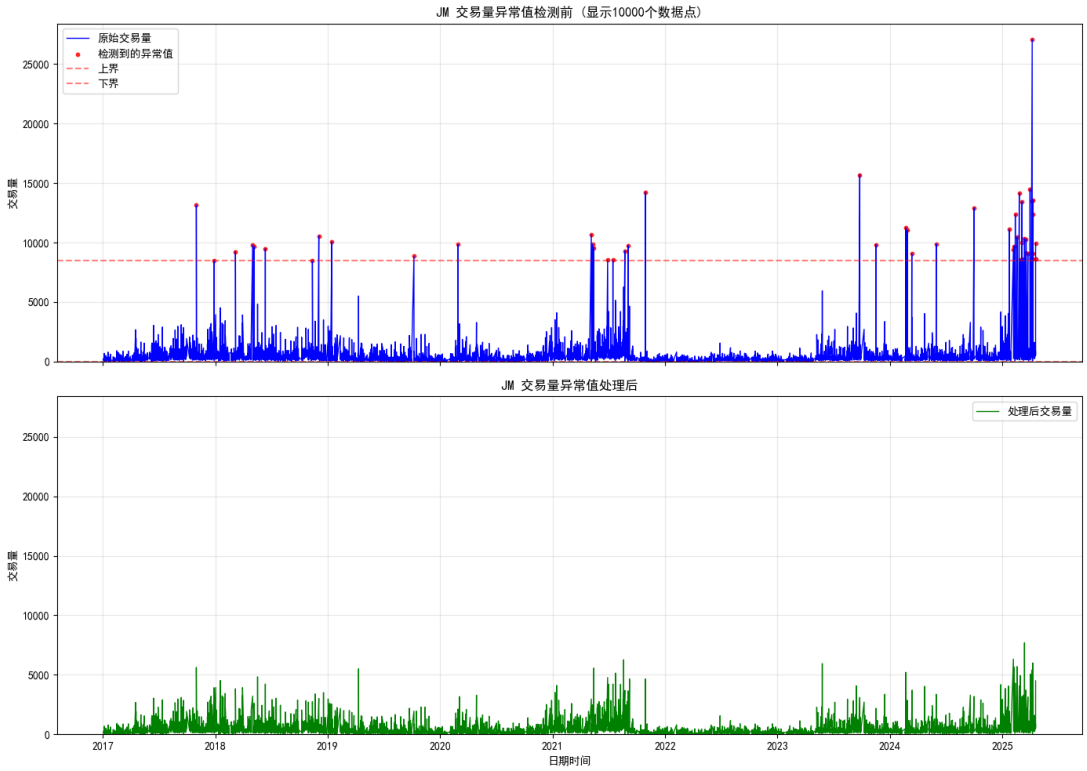
3. 给出异常值处理后的 volume 对比图:



异常值处理后 HC 的 volume 对比图



异常值处理后 I 的 volume 对比图



异常值处理后 JM 的 volume 对比图

4.2 特征提取

提取和涨跌幅强相关的参数:

1. 交易量随时间的变化率
2. 交易量 volume 滞后 30 分钟和滞后 1 天的特征
3. 交叉特征（波动率 × 交易量）

4.2.1 交易量随时间的变化率

原理： 交易量是市场活跃程度的重要指标。根据道氏理论和量价分析理论，价格变动若伴随交易量显著增长，则趋势更可能持续。交易量突增常伴随着主力资金的进出或情绪突变。

数学表达： 交易量的变化率定义如下：

$$\text{Volume Change Rate}_t = \frac{V_t - V_{t-1}}{V_{t-1}} * 100\%$$

其中 V_t 表示当前时间点的交易量，V_{t-1} 为上一个时间点的交易量。

4.2.2 交易量滞后特征（30 分钟和 1 天）

原理： 滞后交易量可以反映市场在过去某一时刻的活跃程度，有助于捕捉市场的短期记忆效应与行为惯性。30 分钟滞后反映了短周期的交易节奏，1 天滞后则体现了日内波动对次日走势的影响。

数学表达： 令 k 为滞后周期，则有：

$$\text{Lagged Volume}_{t-k} = V_{t-k}$$

常用的周期为 $k = 30\text{min},\ 1\text{d}$。

也可构造其相对变化：

$$\Delta V_{t,k} = V_t - V_{t-k}, \quad \text{或} \quad \frac{V_t - V_{t-k}}{V_{t-k}}$$

4.2.3 交叉特征（波动率 × 交易量）

原理： 波动率和交易量是金融市场中两个重要且互补的指标。波动率衡量价格变动的幅度和频率，反映了市场的不确定性。高波动率通常意味着市场情绪不稳定，价格可能会出现大幅波动。相反，低波动率则表明市场相对平静，价格变动较小。交易量则反映了市场的活跃程度和资金流动情况。高交易量通常伴随着较大的市场情绪波动，表明有大量资金进出市场，可能预示着价格的显著变动。低交易量则表明市场较为冷清，价格变动可能较为温和。

通过将波动率与交易量结合,可以更全面地捕捉市场动态。交叉特征 $Volatility \times Volume$ 可以揭示市场剧烈波动前的征兆。具体来说，当波动率与交易量同时升高时，市场更可能出现大行情。这是因为高波动率和高交易量的结合通常意味着主力资金在市场中积极操作，市场情绪较为极端，从而可能导致价格的大幅波动。

此外，交叉特征还可以帮助识别市场的潜在趋势和拐点。在某些情况下，即使波动率较低，但如果交易量突然升高，也可能预示着即将出现的价格波动。反之，高波动率但低交易量可能意味着市场的不确定性，但缺乏足够的资金流动来驱动价格的显著变动。

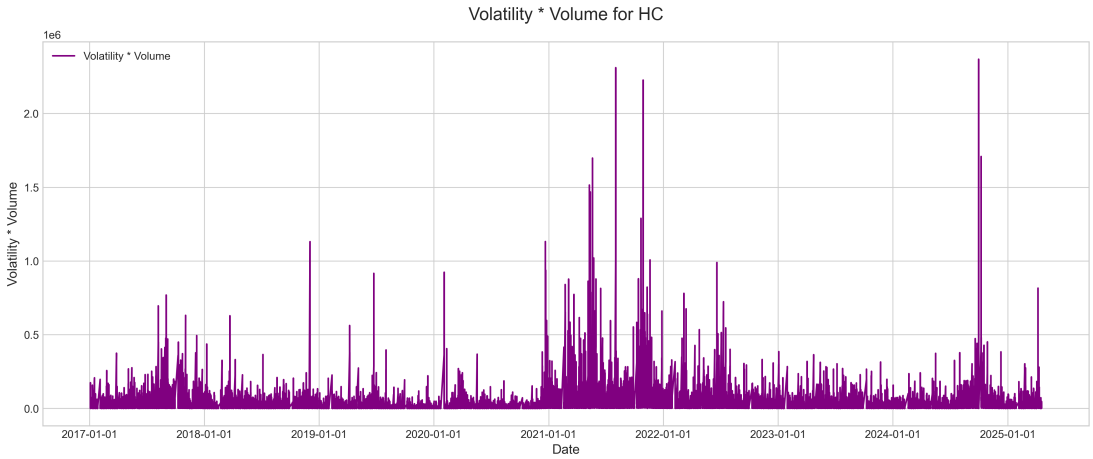
数学表达： 首先定义过去 n 个时间点的波动率为：

$$\text{Volatility}_t = \sqrt{\frac{1}{n} \sum_{i=t-n+1}^{t} (P_i - \bar{P})^2}$$

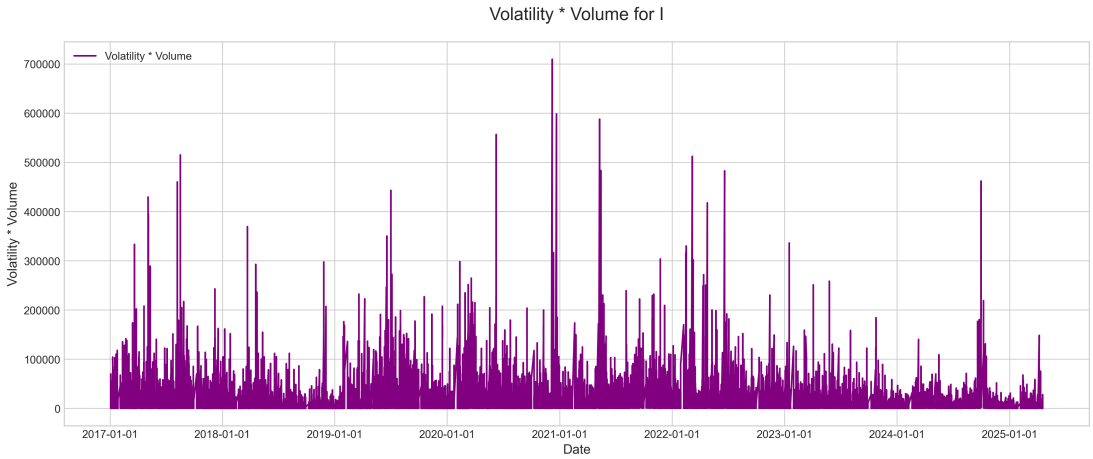
其中 P_i 表示第 i 个时间点的价格，$\bar{P}$ 为该窗口内的平均价格。

交叉特征则为：

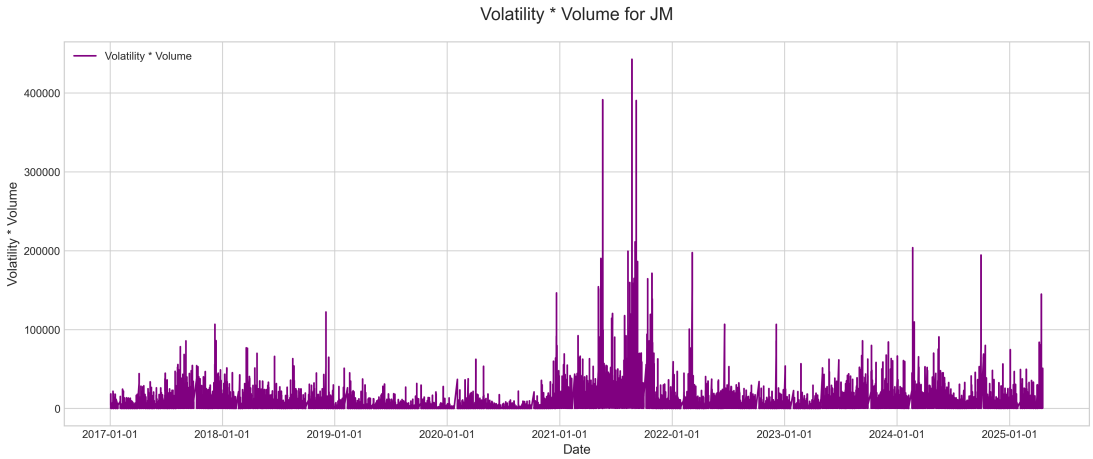
$$\text{Volume-Volatility Interaction}_t = \text{Volatility}_t \times V_t$$



HC 的交叉特征



I 的交叉特征



JM 的交叉特征

4.3 模型选择理由

这是典型的监督学习 + 时间序列回归问题，特点是:

1. 时间依赖性（前后时刻相关）

2. 非线性特征影响（价格、成交量等复杂组合影响未来走势）

LSTM 模型具备以下优势, 利于实现这个任务:
1. 记住较远历史信息
2. 输入序列长度较长
3. 输出依赖时间模式
4. 输入输出为不定长序列

LSTM 模型的数学原理如下:

LSTM 模型与期货涨跌幅预测

门控机制的市场意义

LSTM 的门控机制天然适合捕捉期货市场的三类关键特征：

1: LSTM 门控与市场特征的对应关系

门控类型	数学表达	市场功能
遗忘门	$\mathbf{f}_t = \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{x}_t])$	过滤过时的技术指标 衰减历史波动率影响
输入门	$\mathbf{i}_t = \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{x}_t])$	识别突破性行情 吸收突发新闻事件
输出门	$\mathbf{o}_t = \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{x}_t])$	控制预测信号强度 调节风险暴露程度

期货特征工程

输入特征设计为 5 维向量：

$$\mathbf{x}_t = \begin{bmatrix} \frac{p_t - p_{t-5}}{p_{t-5}} & \text{(5 分钟收益率)} \\ \frac{\text{std}(p_{t-30:t})}{\text{mean}(p_{t-30:t})} & \text{(波动率)} \\ \log(v_t / \bar{v}_{t-60}) & \text{(成交量偏离)} \\ oi_t - oi_{t-30} & \text{(持仓量变化)} \\ \mathbb{I}_{\text{夜盘时段}} & \text{(交易时段标记)} \end{bmatrix}$$

梯度传播的市场解释

$$\frac{\partial \mathcal{L}}{\partial \mathbf{C}_{t-1}} = \mathbf{f}_t + \frac{\partial}{\partial \mathbf{C}_{t-1}}(\mathbf{i}_t \odot \tilde{\mathbf{C}}_t)$$

- 趋势市中 $\mathbf{f}_t \approx 1$
- 震荡市中 $\mathbf{i}_t$ 主导更新
- 极端行情时梯度爆炸抑制

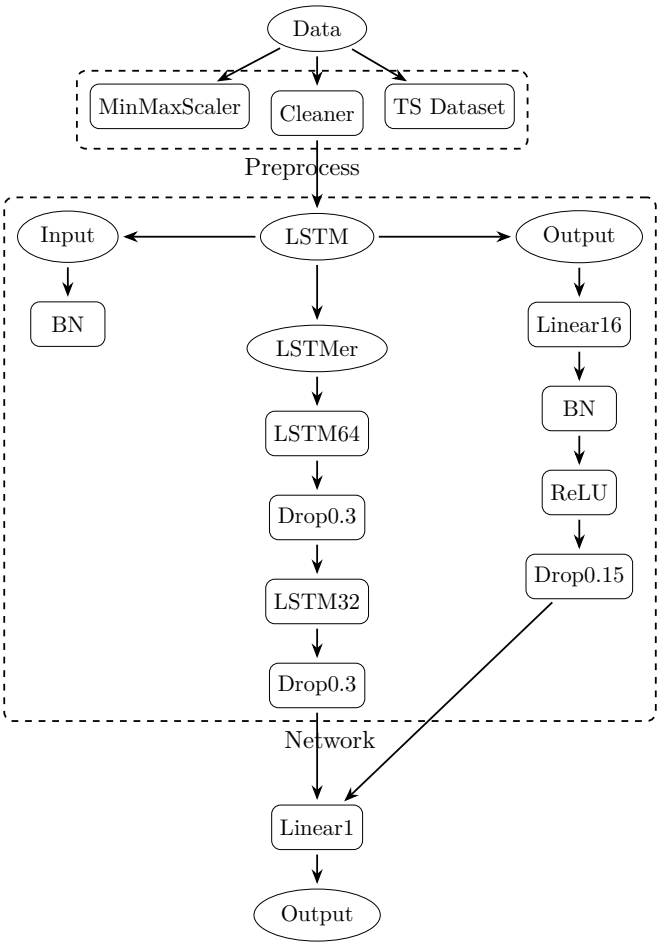
改进的 Peephole 结构

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_f[\mathbf{C}_{t-1}, \Delta p_{t-1}] + \mathbf{b}_f\right)$$
$$\mathbf{o}_t = \sigma\left(\mathbf{W}_o[\mathbf{C}_t, \mathrm{VIX}_t] + \mathbf{b}_o\right)$$

- 价格加速度 Δp_{t-1} 增强趋势判断
- VIX 指数调节风险控制强度

4.4 模型具体实现

模型实现结构图全览:



模型实现结构分层次剖析:

4.4.1 数据预处理

1. **数据标准化**

采用 Min-Max 标准化处理原始期货数据：

$$x_{\text{scaled}} = \frac{x - \min(X)}{\max(X) - \min(X)} \quad \text{(将特征缩放到 [0,1] 区间)} \tag{1}$$

2. **数据清洗**

– 处理缺失值：前向填充（Forward Fill）

– 异常值处理：剔除 $\pm 3\sigma$ 外的价格数据

– 跳空修复：对隔夜缺口进行线性插值

3. **时序数据集构建**
   构建监督学习格式的时序数据：

$$\mathcal{D} = \{(\mathbf{X}_t, y_t) \mid \mathbf{X}_t = [\mathbf{x}_{t-T}, ..., \mathbf{x}_t],\ y_t = p_{t+30}\} \tag{2}$$

其中 $T = 120$ 表示 2 小时历史窗口（对应图中 TimeSeries Dataset）

4.4.2 网络架构

- **输入特征**：包含交易量变化率，交易量滞后特征，交叉特征 3 个维度，经标准化处理后输入网络
- **核心组件**：
  - **BatchNorm 层**：对输入特征进行归一化，设置 $\epsilon = 10^{-5}$ 防止数值不稳定
  - **双层 LSTM 结构**：
    * 第一层：64 维隐藏状态，捕捉短期波动模式
    * 第二层：32 维隐藏状态，提取高阶时序特征
  - **Dropout 层**：
    * 第一层丢弃率 0.3，缓解市场噪声影响
    * 第二层丢弃率 0.15，保留有效特征
- **输出层设计**：
  - 通过 16 维全连接层压缩特征
  - ReLU 激活保证预测收益率非负
  - L2 正则化（$\lambda = 0.01$）控制模型复杂度

4.5 模型的训练与验证

4.5.1 数据预处理流程

- **异常值处理**：
  - 采用前向填充与线性插值组合策略：

$$x_t^{\text{filled}} = \begin{cases} x_{t-1} & \text{单点缺失} \\ \text{linear}(x_{t-k}, x_{t+m}) & \text{连续缺失} \end{cases}$$

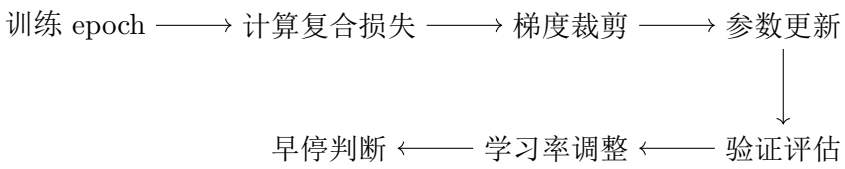
  - 剔除 $\pm 3\sigma$ 外的极端值，保留市场正常波动范围
- **特征工程**：

– 动态缩放：对每个特征列独立进行 MinMax 标准化

$$x^{(j)} \leftarrow \frac{x^{(j)} - \min(X^{(j)})}{\max(X^{(j)}) - \min(X^{(j)})}$$

– 状态保存：持久化 scaler 参数供生产环境复用

4.5.2 优化训练策略



1: 训练循环控制流

关键组件：

- **自适应优化器**：采用 AdamW（改进版 Adam）：
  - $\beta_1 = 0.9,\ \beta_2 = 0.999$ 平衡短期动量与长期方差
  - 解耦权重衰减实现更稳定的 L2 正则
- **动态学习率**：ReduceLROnPlateau 调度器：

$$\eta \leftarrow \begin{cases} 0.5\eta & \text{Val MSE 连续 5 轮 } \nrightarrow \\ \eta & \text{otherwise} \end{cases}$$

- **损失函数**：

$$\mathcal{L} = \underbrace{\frac{1}{N}\sum(y - \hat{y})^2}_{\text{MSE}} + \lambda \underbrace{\|\mathbf{W}\|_2^2}_{\text{L2}} + \alpha \underbrace{\|\mathbf{h}_t - \mathbf{h}_{t-1}\|_2^2}_{\text{状态平滑}}$$

- **正则化设计**：采用多层次正则化策略提升模型鲁棒性，具体实现如下：

类型	实现方式	金融逻辑
L2 权重衰减	Adam 优化器 weight_decay=0.01	抑制市场噪声导致的过拟合
Dropout	LSTM1:0.3 → LSTM2:0.15	渐进式特征选择
梯度裁剪	$\|\nabla W\|_2 \leq 1.0$	防范极端行情梯度爆炸
BatchNorm	输入层标准化	统一量纲
LayerNorm	隐层状态归一化	稳定时序特征

2: 正则化策略配置表

关键技术细节：

- **L2 正则化：**

$$\mathcal{L}_{reg} = \lambda \sum w_i^2 \quad (\lambda = 0.01)$$

通过 AdamW 优化器实现解耦权重衰减

- **Dropout 策略：**
  - 第一层 LSTM 后：0.3 丢弃率过滤噪声
  - 第二层 LSTM 后：0.15 丢弃率保留有效特征
- **梯度约束：**

$$\mathbf{g} \leftarrow \min\left(1.0, \frac{1.0}{\|\mathbf{g}\|_2}\right)\mathbf{g}$$

特别应对交割月合约的剧烈波动

- **归一化处理：**
  - 输入层：BatchNorm1d 处理原始价格
  - 隐层：LayerNorm 适应变长时间序列

- **SMA 数据增强：**
  - **多周期移动平均：**

$$\text{SMA}_{20} = \frac{1}{20}\sum_{i=0}^{19} p_{t-i}$$

$$\text{SMA}_{60} = \frac{1}{60}\sum_{i=0}^{59} p_{t-i}$$

$$\text{EMA}_{10} = 0.18p_t + 0.82\text{EMA}_{10}(p_{t-1})$$

  - **衍生特征构建：**

$$\text{Price-SMA20 Ratio} = \frac{p_t}{\text{SMA}_{20}} - 1$$

$$\text{SMA20-SMA60 Delta} = \text{SMA}_{20} - \text{SMA}_{60}$$

$$\text{EMA10 Slope} = \text{EMA}_{10}(p_t) - \text{EMA}_{10}(p_{t-5})$$

  - **金融逻辑：**
    - 价格/均线比率识别超买超卖状态
    - 双均线差值判断趋势强度
    - EMA 斜率捕捉短期动量变化

4.5.3 验证与模型选择

- **双重评估指标：**

$$\text{MSE} = \frac{1}{N}\sum(y_i - \hat{y}_i)^2 \quad \text{(强调极端误差)}$$

$$\text{MAE} = \frac{1}{N}\sum|y_i - \hat{y}_i| \quad \text{(衡量平均偏差)}$$

- **早停机制：**

  - 基于验证集 MSE 的 patience=10 策略
  - 保留最佳模型状态：选择验证集均方误差最小对应的参数 $\mathbf{W}_{\text{best}}$，即

$$\mathbf{W}_{\text{best}} = \min_{\mathbf{W}} \text{MSE}_{\text{val}}(\mathbf{W})$$

- **交叉验证**：时序分割 (TimeSeriesSplit) 保持时间依赖

4.5.4 实现优化亮点

- **内存管理**：
  - DataLoader 的 pin_memory 加速 GPU 数据传输
  - 梯度累积支持超大 batch size
- **可复现性**：
  - 设置全局随机种子 (seed=42)
  - 确定性算法配置
- **生产部署**：
  - 模型量化 (FP16) 提升推理速度
  - 异常输入自动检测与恢复

4.6 模型预测效果与改进建议

5 源码与文档