1. 特征工程

1.1 特征构建

构建了众多常用的技术指标，包括：

- Average True Range (ATR)
- Ease of Movement (EMV)
- Force Index (FI)
- Money Flow Index (MFI)
- Standard Deviation (std)
- Relative Strength Index (RSI)
- Stochastic Oscillator (KDJ)
- Ultimate Oscillator (UOS)
- Average Directional Index (ADX)

以日内波幅作为 target，计算各个特征与其的相关系数：

Table 1 **部分技术指标与日内波幅的相关系数**

ATR	**EMV**	**EMV_ma**	**FI**	**MFI**	**Std**	**RSI**
0.5419	-0.38	-0.301	-0.1682	-0.1467	0.43474	-0.1996
KDJ_k	**KDJ_d**	**KDJ_j**	**UOS**	**ADX_+di**	**ADX_-di**	**ADX**
-0.1272	-0.10728	-0.07414	0.178836	0.02433	0.11329	-0.16211

下图为各个特征之间的相关系数热图矩阵：

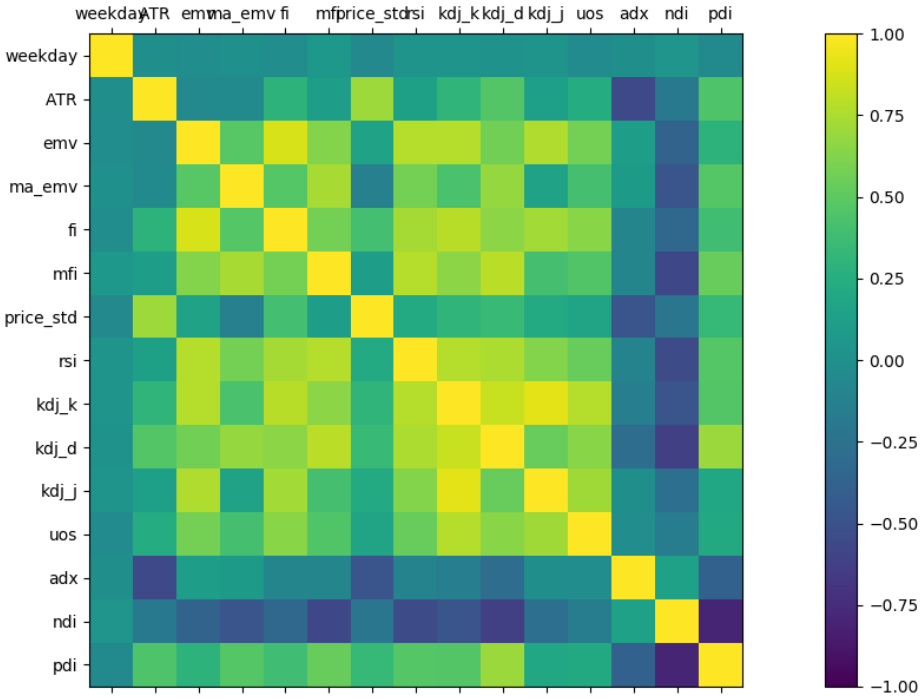

Figure 1 **特征的相关系数矩阵热图**

另外，尝试构建了一些新特征。

1) True Range Day (TRd)

据观察，日内上午和下午通常具有不同的走势，因此分别计算日内上午(TRm)和下午的波幅(TRa)。
与波幅的相关系数：TRm: 0.765，TRa: 0.774。

2) Volume Weighted ATR (VWATR)

计算了每日交易量与波幅的相关系数，为 0.5763。交易量与波幅相关性较高，因此考虑结合交易量与价格构造新的特征。尝试构造交易量加权的 ATR 指标，设N为窗口大小，则第t日的指标值为：

$$VWATR_t = \frac{\sum_{k=1}^{N} ATR_{t-k} \times Vol_{t-k}}{\sum_{k=1}^{N} Vol_{t-k}}$$

即窗口内的交易量加权的 ATR 均值。
与波幅的相关系数为 0.6598。

3) Volume Std (Volstd)

简单的窗口内交易量标准差，仿照窗口价格标准差计算得来。
与波幅的相关系数为 0.2719。

4) Cumulative Absolute Change (CAC)

考虑到有时两个波幅相同的交易日，其日内 tick 级数据的走势特征可能迥然不同。为了综合考虑这种日内的波动形式对波幅的影响，尝试构建了日内累-计波动值特征。
累积价格波动值：

$$CAC_{price} = \sum_{t} abs(Price_t - Price_{t-1})$$

累计交易量波动值：

$$CAC_{vol} = \sum_{t} abs(Vol_t - Vol_{t-1})$$

其中 t 为日内的第 t 分钟交易的 tick。
与波幅的相关系数：CAC_Price: 0.5625，CAC_Vol: 0.4649。

1.2 特征选择

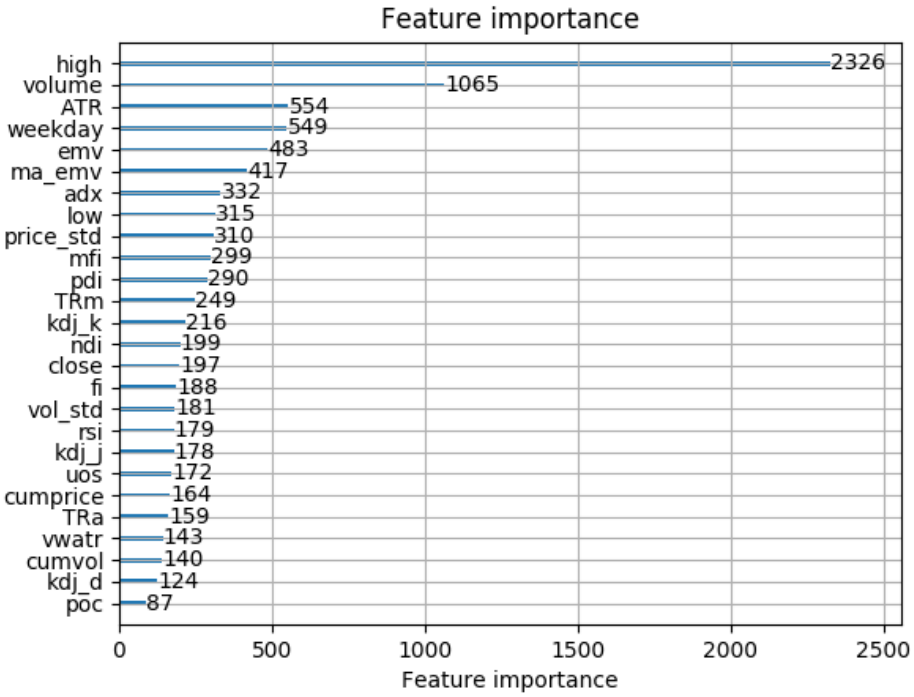

Figure 2 由 LightGBM 模型生成的特征重要度排名

根据 lightgbm 的 plot_importance 结果，选择了特征 high, volume, weekday, ATR & EMV。

2. 训练与预测

选择了 sid 为 000001 的股票从 2016 年 1 月 1 日至 6 月 1 日的 tick 和 bar 数据，使用 LSTM 进行建模训练并预测。

2.1 timesteps 的设置

Timesteps 作为 LSTM 区别于全连接 DNN 特有的参数，对模型的泛化能力具有重要的影响。尝试了多种不同的 timesteps 在 train 集上的拟合结果与 test 集上的预测表现，结果如下：

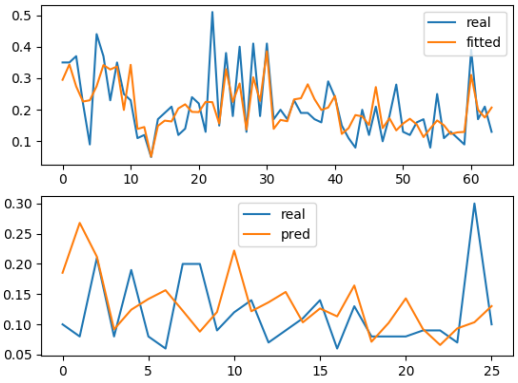

Figure 3 Timesteps=1 时的拟合结果

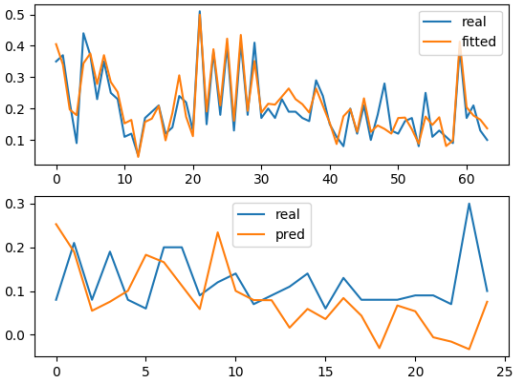

Figure 4 Timesteps=2 时的拟合结果

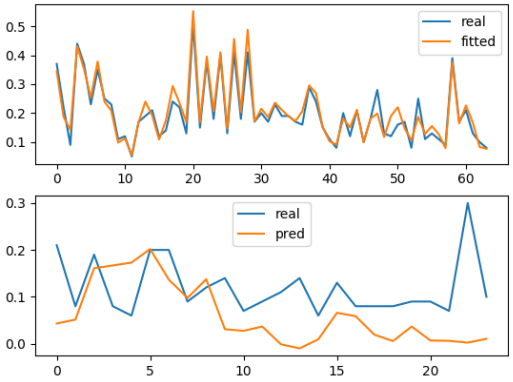

Figure 5 Timesteps=3 时的拟合结果

当 timesteps 取的越大，模型在训练集上的拟合越紧密，而在测试集上表现则越差，甚至会出现负值的预测结果。因此选择 timesteps=1 在这个问题下模型表现最好。

2.2 加入 target 的滞后值作为特征

检验如果加入 target 作为训练特征是否能提高泛化能力，即使用昨日的 target 作为今日的特征。

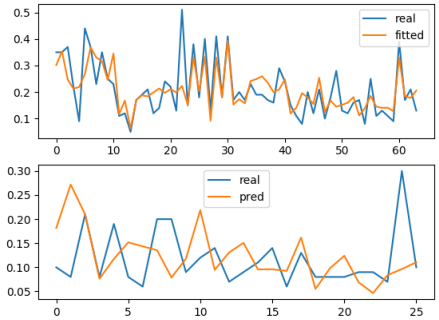

Figure 6 加入 target 滞后值

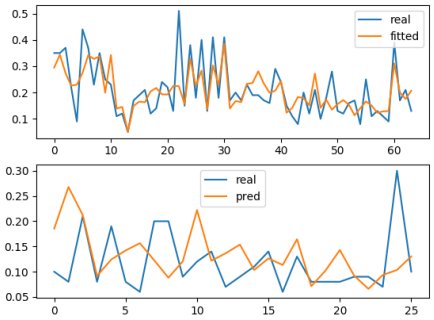

Figure 7 不加入 target 滞后值

二者相差不大，秉持奥卡姆剃刀原则，可以不加入 target 滞后值作为特征。