Volatility Predictions

May 2024

Data description and cleaning

The dataset we want to analyse is that of stock price data for 6 different stocks, sampled at 1 minute frequency, on the trading days from Jan. 2nd, 2017 to Dec. 29th, 2017. We have trading days beginning at trading time 9:30 am to 4:00 pm (391 minutes in total). After looking closely at the data, we realize that there are several abnormalities that we either modify or keep in mind for future analysis:

- Stock a and d both contain 93 irregular price values, (0.0 for stock a and 1.0 for stock d), we replace these prices by the prices one-minute prior to them.
- Day 327 in the dataset, Nov. 24th, 2017, is a half trading day and contains only 211 trading minutes. We delete this day for all 6 stocks.
- Stock a, c, d, f contain 71, 31, 18, 1371 missing values, respectively. We fill these NaN values by their most recent available prices (indicating a zero volatility over this period). If the missing period is at the beginning of a trading day (starts from 9:30 am), we fill these NaN values by their next available prices.
- As indicated by the price time series plot in Figure 2 in the appendix, stock c experienced a 45 % price drop at the opening minute on day 149. We delete this day for stock c.

Warm up: Linear regression on the previous two months

For some stocks, there is rapid daily variation (stock 'd' for instance), presumably from algorithmic high frequency trading, that does not reflect the overall trends. We will assume this could be separately modeled if needed, and so we remove this effect by smoothing with a Gaussian kernel with a width of one minute (see, e.g, Bezerra et al.(2017)). In fact, Ghysels et al.(2002) found that volatility predictions are not greatly enhanced by using high frequency sampling.

Therefore, in this section, we will simplify the problem to one regarding daily opening or closing prices. Since opening and closing prices are well correlated, we will content ourselves with closing prices as it is often done in the literature. On the other hand, Alizadeh et al.(2006) considered daily ranges as a tool for volatility forecasting, Ghysels et al.(2002) confirmed that the use of daily ranges and absolute returns are powerful for volatility forecasting, especially in autoregressive methods. Thus, we will examine the daily price ranges as well in this project.

Our goal is to forecast the volatility over the next month following the end of the samples for all 6 stocks. Consider the price sequence as a stochastic process, for any time interval $[t-N, t]$, the empirical estimation of the quadratic variation of a stochastic process is given by the following:

$$\sigma_N(t)^2 = \frac{1}{N+1} \sum_{i=t-N}^{t} (r_{i,i-1} - \overline{r_{t-N,t}})^2 \tag{1}$$

Where $r_{i,i-1}$ is the return of the stock on day i, whereas $N = 21$ is the number of trading days in a month. $r_{i,i-1} = log(P_i/P_{i-1})$ We include the daily volatility time series plot for all 6 stocks in Figure (??) in the appendix.

Our first attempt will be a linear regression model, where the inputs features are either:

- Historical volatility over the previous two months ($2 * N$ trading days) following the result of (e.g., Ghysels et al.(2006)) that volatility prediction is fairly captured by the past two months:

$$\sigma(t+h) = \beta_0 + \sum_{i=t-2N}^{t} \beta_i \sigma(i) \tag{2}$$

- Daily price ranges for the past two months:

$$\sigma(t+h) = \beta_0 + \sum_{i=t-2N}^{t} \beta_i [hi - lo](i) \tag{3}$$

- A combination of both of the above regressors.

,

The implied volatility will then be estimated for the following month. A quick comparison of MSE shows the superiority of the first model. We keep the data of the last month in the vault and use cross validation on the rest of the data, the resulting predictions for the last month for the first model along with the annualized volatility for the next month are shown in the appendix. Although for stock "a" the predictions seem to fit for the the majority of the

month, the model doesn't do a good job capturing the volatility change, probably because of the dependence of volatility on the most recent price changes, and the fact that for some stocks the volatility was quite bursty at certain earlier periods, we should also question whether intraday volatility should be included.

Model Description

The empirical estimation of the quadratic variation of a stochastic process is given by the following realized variance, $RV_{t+1} = \sum_i r^2_{i,i-1}$. Here we consider the partition of time interval $[t, t+1]$ into $1/391$ equal sub-intervals. For this project, t and $t+1$ can be seen as the beginning and closing minutes for day t. Therefore, $\sqrt{RV_t}$ measures the daily volatility on day t, including the intraday volatility (figure 5).

Almon's approach to modelling distributed lags has been used very effectively more recently in the estimation of the so-called MIDAS model. The MIDAS model (developed by Eric Ghysels and his colleagues - e.g., see e.g., Ghysels et al.(2004)) is designed to handle regression analysis using data with different observation frequencies. The acronym, "MIDAS", stands for "Mixed-Data Sampling". We consider here a regression model with polynomial coefficients in the lags. Specifically,

$$RV_{t,h} = \beta_0 + \sum_{k=1}^{k_{max}} \beta(k, a_0, a_1, a_2) RV_{t-k+1}. \tag{4}$$

where $RV_{t,h}$ is the averaged daily realized variance over the period $[t, t+h]$. We parameterize the coefficients as follows, $\beta(k, a_0, a_1, a_2) = a_0 + a_1 k + a_2 k^2$. The main idea of MIDAS regression is to use regressors which may have different frequency from the response variable.
On the other hand, based on HARCH, (Corsi .(2006)) developed the HAR-RV model, a prediction model that captures the long range dependencies of the data by incorporating lagged realized volatility over different time horizons:

$$RV_{t,h} = \beta_0 + \beta_d RV_t + \beta_w RV_{t-5,5} + \beta_m RV_{t-21,21} \tag{5}$$

Therefore, the model predicts future volatility using a daily, a weekly and a monthly component, respectively. The HAR-RV model can be seen as a prediction which uses the exponential smoothing of lagged values of RV_t.

Model selection and reporting the results

In this section, we perform the model fitting and selection on all 6 stocks, using the models mentioned above. Our setups are summarized as follows,

1. For Almong Lag regression, we set $k_{max} = 9$ and use the past 10 day volatility information

2. Each stock has approximately 210 samples to work with. We split the dataset into two part, the training set contains the first 70% of the data and is used to fit different models. We evaluate the performances of different models on the rest part of the data, conduct model selections and estimate the standard deviations for the models' predictions. to forecast the future monthly volatility. The performance of both models on the test data is plotted in the appendix.

3. Both models are fitted by least squares and are performed by *scikit-learn*'s *LinearRegression* in Python. In general for MIDAS, we could have used *minimize* from *scikit.optimize* to find the parameters coming from non-linear regression, but we turned our model into a linear one with 4 coefficients, so we didn't have to.

4. Looking at the the plots of the prices for the six stocks, it seems that some stocks have varying behaviours at different periods of the year, stock 'e' for instance has a big burst in the first half of the year, other stocks exhibit similar behaviours too. It is wise to think that the second half of the year is more indicative of volatility for some of these stocks, we check how these models perform when only trained on the second half of the year. *HAR-RV* does better on the truncated dataset for stocks: 'd', 'e' and 'f', not so much for the other stocks (figures 8 and 9).

We now report the best model for each stock, the performance of the model, the prediction for the next month volatility, and the fitted parameters.

a) HAR-RV is best: Coefficients for stock A: β_0: 0.01, β_1: 0.05, β_5: 0.12, β_{21}: 0.24 . Predicted Future Volatility for stock A after the samples: 0.0160. 95% Confidence Interval: $(0.0155, 0.0166)$.

b) Almon/MIDAS is best: Coefficients for stock B: [-0.04, 0.55, -0.22, 0.02] Estimated future volatility for stock B at the end of the samples: 0.31755802240398323 95% confidence interval for stock B:$(0.30, 0.33)$

c) Almon/MIDAS is best: Coefficients for stock C: [0.0088388 0.24399423 -0.07324876 0.00580492] Estimated future volatility for stock C at the end of the samples: 0.024773475394830675. 95% confidence interval for stock C: $(0.024773475394830505, 0.024773475394830845)$

d) HAR-RV truncated is best: Coefficients for stock D: β_0: 4.04562959e-01, β_1: 0.0017297805222307213, β_5: 0.08045970571702661, β_{21}: -0.6960134479514177. Predicted Future Volatility for stock D after the samples: Prediction: 0.3588. 95% Confidence Interval: $(0.3573, 0.3602)$

e) HAR-RV truncated is best: Coefficients for stock E: β_0: 0.01976635 β_1: 0.005, β_5: 0.187, β_{21}: -1.250. Predicted Future Volatility for stock E after the samples: 0.0205. 95% Confidence Interval: $(0.0204, 0.0206)$

f) HAR-RV trundacted is best: Coefficients for stock F: β_0: 1.62830011e-02, β_1: -0.012, β_5: -0.067, β_{21}: -0.491. Predicted Future Volatility for stock F after the samples: 0.0254. 95% Confidence Interval: $(0.0245, 0.0265)$

Appendix

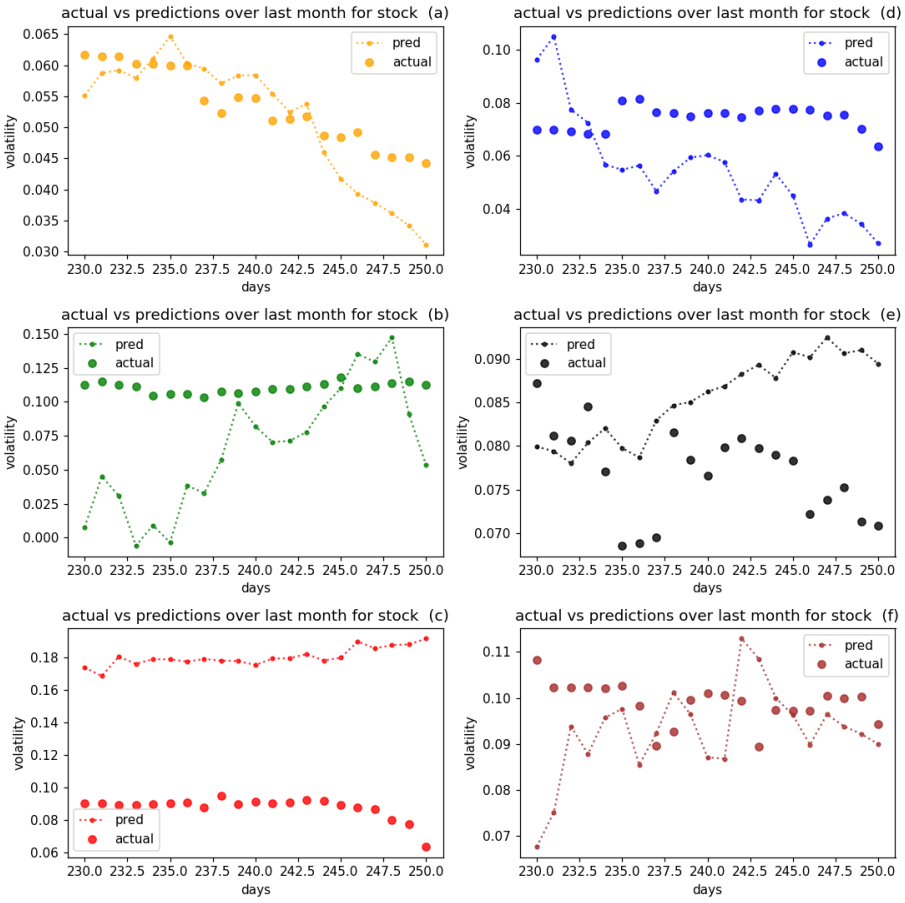


Figure 1: Naive Linear regression: Volatility predictions versus actual values for the last month

Stock	Estimation and confidence intervals
'a'	1.6% +/- 3.0%
'b'	21.0% +/- 2.3%
'c'	72.7% +/- 0.3%
'd'	24.8% +/- 10.9%
'e'	29.2% +/- 2.1%
'f'	26.1% +/- 3.0%

Table 1: Prediction for next month using naive linear regression model

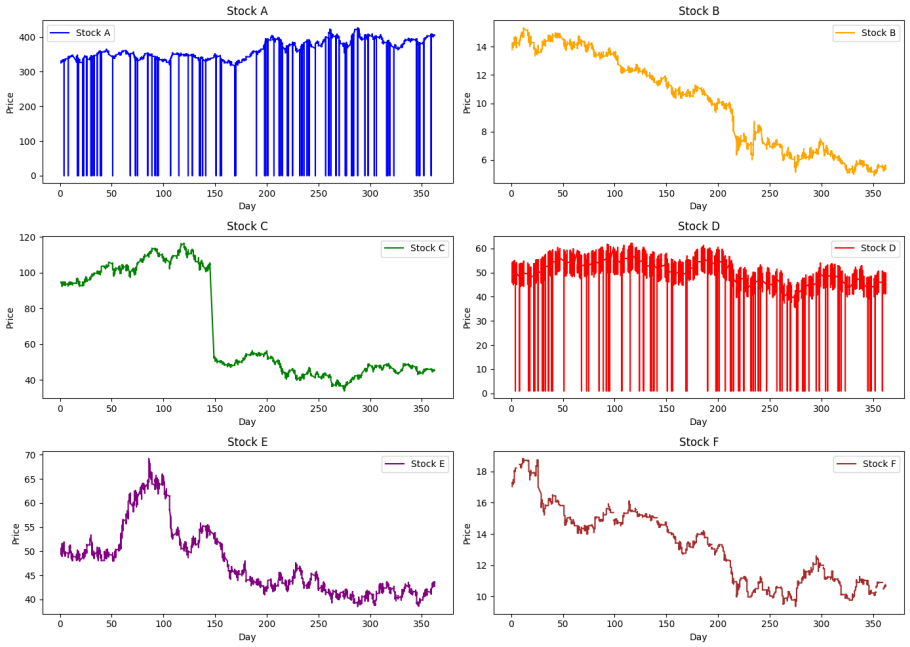


Figure 2: Raw prices of all 6 stocks

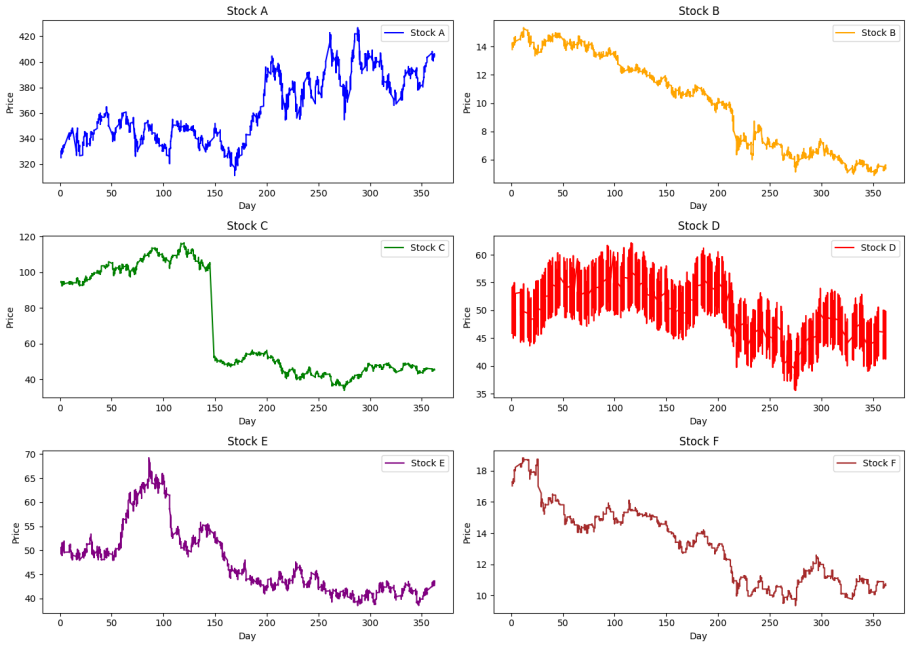


Figure 3: Prices of all 6 stocks after initial cleaning

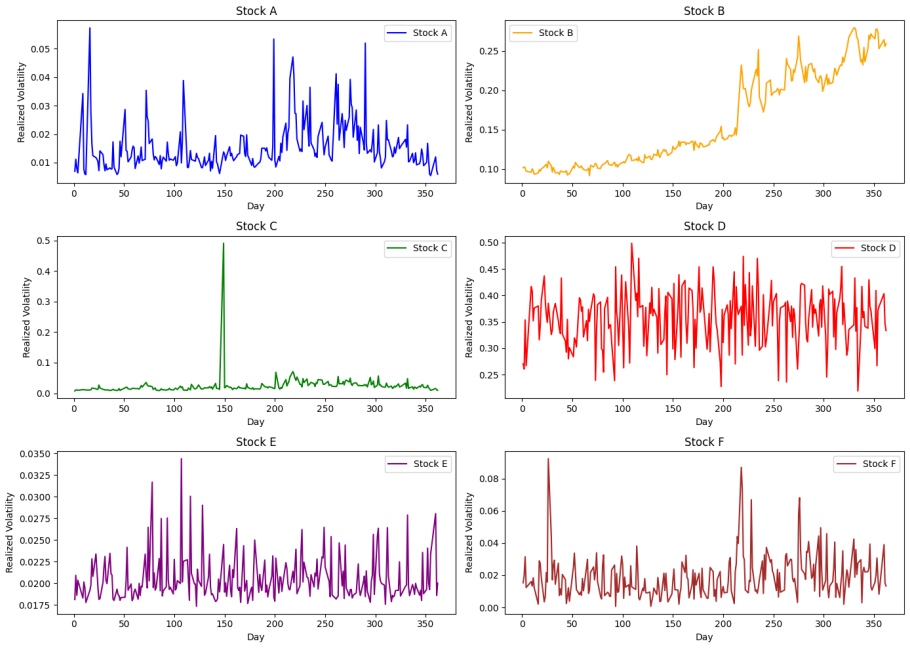


Figure 4: Realized volatility for all 6 stocks

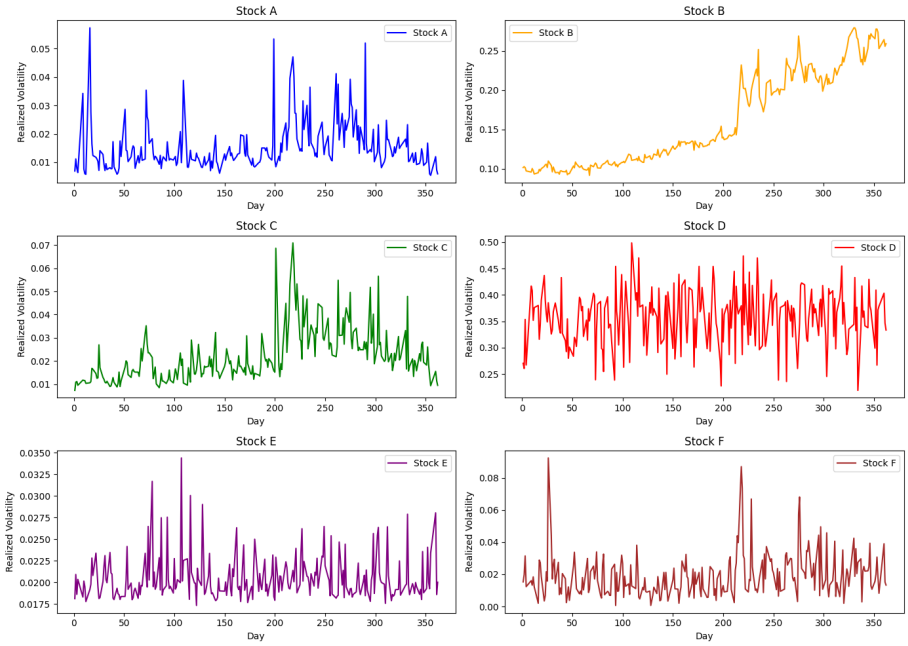


Figure 5: Realized volatility for all 6 stocks after we deleted the jump day

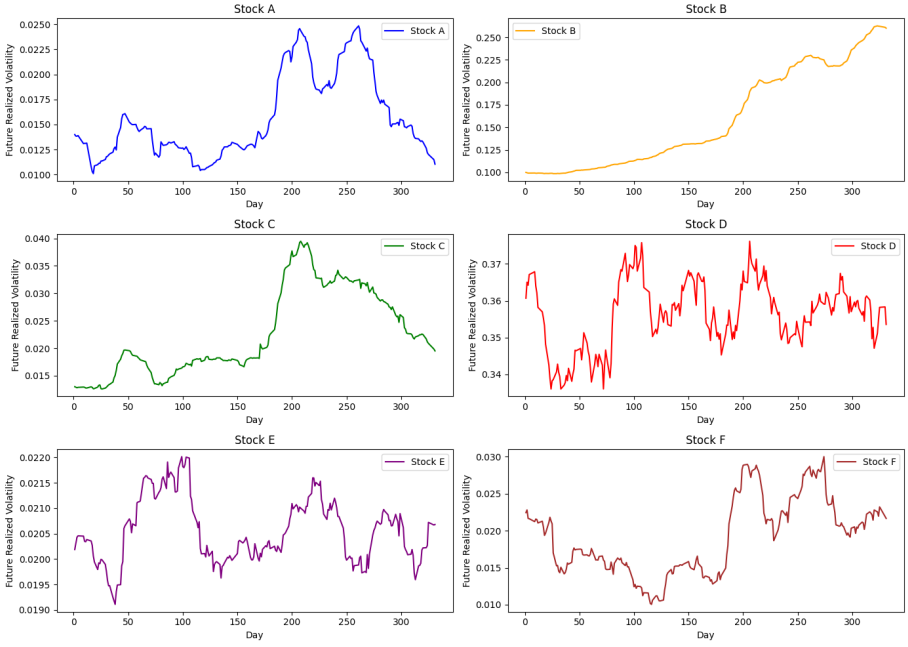


Figure 6: Future volatility over next month for each stock

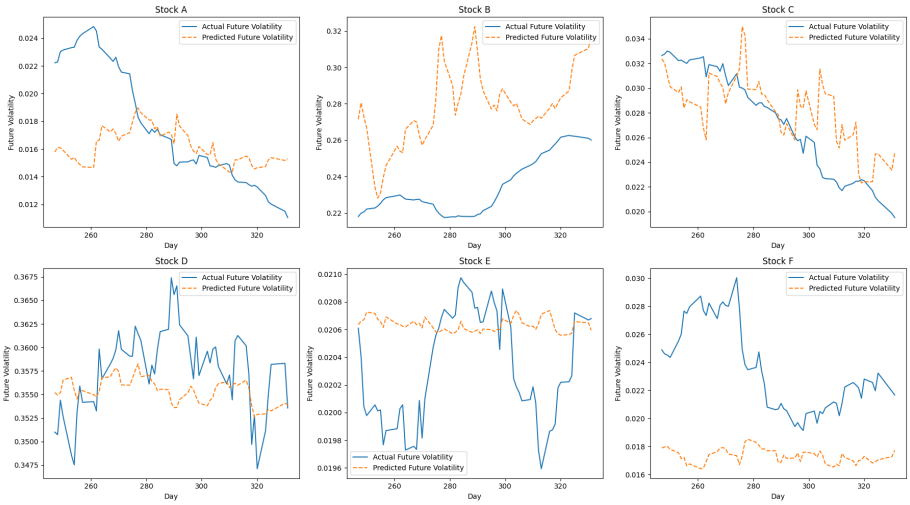


Figure 7: Almon/MIDAS performance on the test set

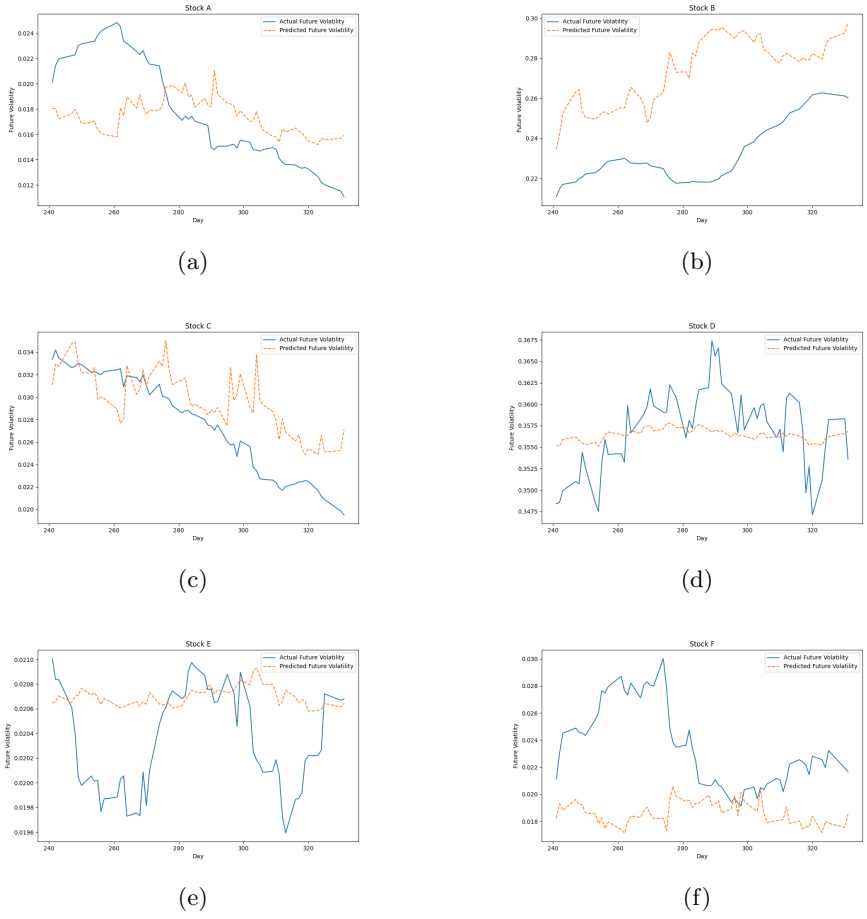
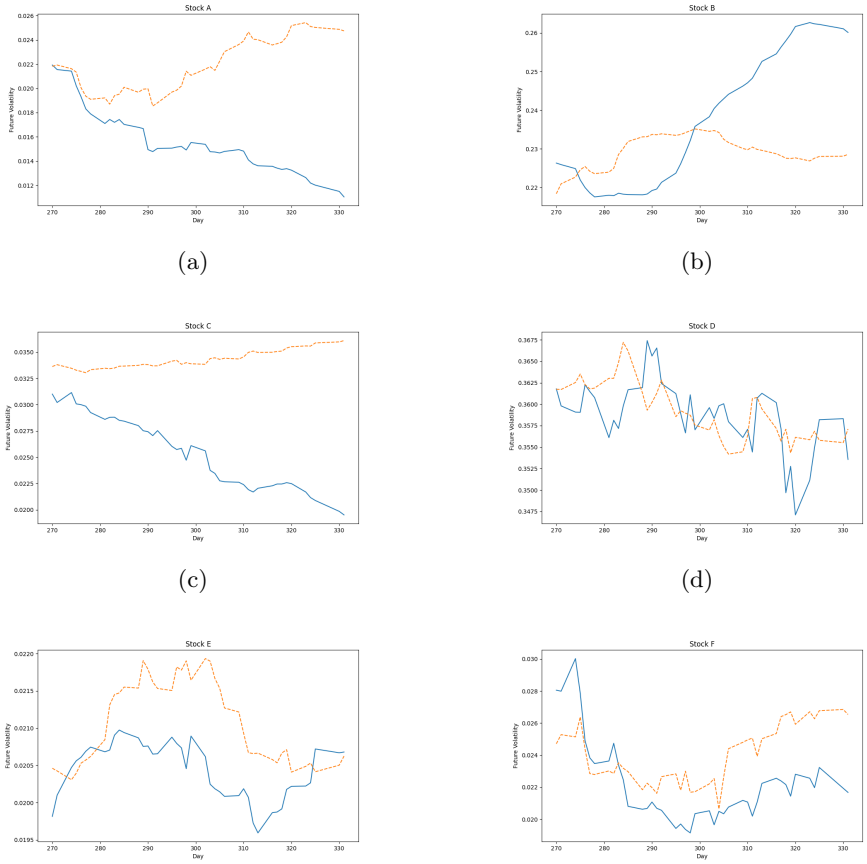


Figure 8: *HAR-RV* performance on the test set



(a) (b) (c) (d) (e) (f)

Figure 9: *HAR-RV* performance on the test set after truncation