Erasmus University Rotterdam

Erasmus School of Economics

Bachelor Thesis Econometrics and Operations Research

A Tree-Based Machine Learning Approach to Forecasting Volatility

Stijn Koene (585185sk)

Supervisor:	dr. Onno Kleen
Second assessor:	M.F.O. Welz
Date final version:	2nd July 2023

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

We inspect which model is most accurate in predicting the realised variance of stocks, by comparing 15 different forecast series generated by four sets of models. The models we use are: four extensions to the heterogeneous autoregressive (HAR) model, five shrinkage models, two tree-based machine learning (ML) models and three sets of combined forecasts. The regular HAR model is used as benchmark. To perform our research, we examine data from 24 companies included in the Dow Jones Industrial Average. We conclude that among the ML models random forests generate the highest reduction in MSE, whereas among all models the forecast combinations perform best. Moreover, we aim to add interpretation to the ML models by utilising the Shapley additive explanation values. From this we conclude that the lagged values of the realised variance are the most important features for forecasting volatility.

1 Introduction

Stock volatility plays a central role in the decision-making process of market participants, from individual investors to institutional traders. Accurate forecasting of volatility is hence crucial for risk management, portfolio optimisation and the pricing of derivatives. This research delves into which models, or combination of models, are best at producing accurate realised variance forecasts. Volatility forecasting is a widely discussed topic in the academic literature and goes back decades (Christoffersen and Diebold, 2000; Engle and Patton, 2001; Andersen et al., 2003). A prominent class of models is the family of ARCH models, first proposed by Engle (1982) and Bollerslev (1986). ARCH models introduced the concept of time-varying volatility by incorporating past information. These models paved the way for subsequent advances and laid the foundation for more complex approaches to volatility forecasting.

Later on, Corsi (2009) introduced the heterogeneous autoregressive (HAR) model, which has grown to become a second widely used family of models for forecasting volatility. Corsi (2009) argued that the ARCH models were inadequate to capture the long-memory effects of volatility, whereas the HAR model "combines nonparametric realised variance measured at different frequencies with a parametric autoregressive model" to capture the long-memory of volatility, according to Christensen et al. (2021). Eventually the HAR model took over the title of "the benchmark" model in the academic literature (Hansen and Lunde, 2005), previously held by the GARCH(1,1) model of Bollerslev (1986). This naturally led to several extensions of the HAR model (Corsi and Renò, 2012; Patton and Sheppard, 2015), from which we include four in our paper.

In recent years, advances in computational techniques and access to vast amounts of financial data have led to the development of more sophisticated volatility forecasting models. These models incorporate various statistical and econometric methods as well as machine learning and artificial intelligence algorithms (Christensen et al., 2021; Bucci, 2020). Such advances have enabled researchers and practitioners to improve the accuracy and reliability of volatility forecasting. However, there is not much academic literature that evaluates multiple machine learning (ML) methods for volatility forecasting; there is mainly literature that compares a single ML model with benchmark models. Some examples of available literature in this category are Luong and Dokuchaev (2018) and Khaidem et al. (2016) who investigate random forests (RF); Audrino and Knaus (2016), Caporin and Poli (2017) and Li et al. (2022) use lasso (LA); Bucci (2020), Lei et al. (2021) and Rahimikia and Poon (2020) apply neural networks (NN). As an exception, Christensen et al. (2021) compare several ML methods in their research, so we aim to extend their work in our research by adding another set of generated forecasts.

The purpose of our research is to investigate whether ML models, or their combined forecasts, significantly outperform the benchmark HAR models in forecasting stock volatility. Moreover, we aim to add interpretation to the ML methods in our research in order to uncover the 'black box' surrounding ML models.

We examine our research questions by evaluating five HAR-type models and comparing these models with seven ML models and three model combinations. The dataset we use in our research includes five macroeconomic- and seven firm-specific variables collected between January 3, 2000 and December 31, 2021. More specifics on the variables used can be found in Appendix A. The HAR models we use in our paper are: the standard and log transformed HAR model from Corsi (2009), the semivariance HAR (SHAR) model from Patton and Sheppard (2015) and the HARQ model from Bollerslev et al. (2016). Furthermore, we use the following ML models: Lasso (LA), Ridge Regression (RR), Elastic Net Regression (EN), Adaptive Lasso (A-LA), Adaptive Elastic Net Regression (A-NE), Random Forest (RF) and Extreme Gradient Boosting (XGB). Finally, three forecasting combination models are used: a combination of all HAR models, a combination of all ML models and a combination of all models.

For uncovering the 'black box' around ML models, we use the Shapley additive explanation (SHAP) values of Lundberg and Lee (2017). SHAP values are a commonly used additive feature attribution method in the context of ML model interpretation (Medeiros et al., 2021; Lim et al., 2021). The SHAP score is a generalisation of the Shapley value, which determines the contribution of each input feature to the model's prediction. This contributes to the current academic literature as the 'black box' around ML methods is often criticised. By including methods of interpretation, this 'black box' can be further exposed. The use of SHAP is a contribution to the literature as most papers on volatility forecasting only use Accumulated Local Effects (ALE) for interpreting ML models. Moreover, we use the Model Confidence Set of Hansen et al. (2011) to further interpret our results.

The findings in this paper can be categorized in two parts. First, the combined forecasts obtain the highest reduction in mean squared error (MSE) relative to the regular HAR model, where the highest reduction is seen for the combined forecasts of all models with a reduction of 10.3%. Moreover, according to the Diebold-Mariano test, these combined forecasts significantly reduce the MSE for more than 50% of the stocks studied, compared to three extensions of the HAR model, at a 10% significance level. Additionally, the shrinkage models are significantly outperformed by the combined forecasts. Moreover, the forecasts obtained from the RF model also show significant reductions in the MSE, making this model most compatible with the forecast combinations. Second, we find that the lagged values of the realised variance play the most important role in forecasting volatility, according to their SHAP values. This could be used to argue for the use of HAR models, and may be the reason why the RF model does not perform significantly better than the regular HAR model.

Our work contributes to the existing academic literature in the following three ways. First, we extend the dataset used in Christensen et al. (2021) by adding the positive and negative semivariances of volatility as described in Section 2. Additionally, the length of the dataset is extended by three years. Second, as mentioned above, this paper adds the ML interpretation method of Lundberg and Lee (2017), SHAP values. SHAP values are not extensively studied in volatility forecasting, as Accumulated Local Effects (ALE) are mostly used to add interpretation to ML in this field of academic research (Christensen et al., 2021; Kleen and Tetereva, 2022). Third, this paper contributes by adding a forecast combination of several models, which is a widely studied topic in the current literature (Timmermann, 2006; Newbold and Harvey, 2002), but not for volatility forecasting using ML methods.

The remainder of this paper is structured as follows. Section 2 is an overview of the data. In Section 3 the applied methodology is reviewed. In Section 4, we present the results and conclude the paper in Section 5.

2 Data

This section describes the data used in this paper. It also provides some relevant summary statistics on the data. The decision on which data to include in our research is mainly based on the research by Christensen et al. (2021), with a few exceptions our data mostly coincides with this paper. Furthermore, we use two additional variables suggested by Kleen and Tetereva (2022) in this research to extend the dataset of Christensen et al. (2021).

In this paper we use data from 24 firms included in the Dow Jones Industrial Average (DJIA) index prior to the recomposition on August 31, 2020. The complete list of ticker symbols of the included companies in this research is: AAPL, BA, CAT, CSCO, CVX, DIS, GE, GS, HD, IBM, INTC, JNJ, KO, MCD, MMM, MRK, MSFT, NKE, PFE, RTX, TRV, VZ, WMT and XOM. Six companies are excluded due to limited availability of their data for our sample, these companies are American Express (AXP), Dow Chemical (DOW), JPMorgan Chase (JPM), Procter & Gamble (PG), UnitedHealth Group (UNH) and Visa (V). The dataset used in this paper consists of five macroeconomic variables from the United States and seven firm-specific variables, covering the period from January 3, 2000 to December 31, 2021, giving a total of $T = 5512$ trading days. The variables are collected from various sources, such as the Federal Reserve Economic Data (FRED) database, Yahoo Finance, and a database constructed by Kleen and Tetereva (2022). A general overview of the sources of each variable is given in Appendix A.

The variables are described as follows. First, we look at the six firm-specific variables. The inclusion of the daily, weekly and monthly lags of realised variance is a commonality in almost every academic paper on volatility forecasting, starting from Bollerslev (1986). This is because a significant part of future volatility can be inferred from historical volatility. Hence, these variables are naturally included in our research. The lagged realised variances are denoted by RVD, RVW and RVM respectively. The realised variance for each firm comes from the Kleen and Tetereva (2022) database, from which we compute their lags.

Next, we include two other daily stock-specific variables, the semivariances of the positive and negative returns, denoted by RV^+ and RV^- respectively. These variables are again obtained from the database by Kleen and Tetereva (2022). We use these variables to account for a possible difference in the magnitude of the effect from positive and negative volatility shocks. Furthermore, we include the dollar trading volume of a firm's stock on day t (\$VOL), which we transform by first taking its logarithm and then taking the first difference of the logarithmic transformation. The dollar trading volume is taken from Yahoo Finance. Finally, we include the one-week momentum of a company (M1W), which is calculated as the weekly difference in closing prices of a company's shares, expressed as a percentage. This variable is again obtained from Yahoo Finance.

Compared to Christensen et al. (2021) we use two different variables, we use RV^+ and RV^-, whereas Christensen et al. (2021) use earnings announcement (EA) and implied volatility (IV) in their research. We do not include EA and IV due to time constraints, as obtaining these variables is very time consuming and Christensen et al. (2021) have not published their database publicly.

Next, we use five macroeconomic variables, which can be described as follows. The first variable we use is the Economic Political Uncertainty (EPU) index from Baker et al. (2016). Furthermore, we use the CBOE Volatility Index (VIX), which is a very widely used predictor in volatility forecasting (Bekaert and Hoerova, 2014; Kleen and Tetereva, 2022). Our next variable is the US 3-month Treasury bill market rate (US3M), of which we take the first differences to account for possible non-stationarity. Finally, the Hang Seng stock index (HSI) and the ADS index proposed by Aruoba et al. (2009) are used as predictors. The HSI is transformed as the daily return on the log-squared closing prices. Our macroeconomic variables are fully in line with those used in Christensen et al. (2021).

Table 1 shows a list of all explanatory variables along with some descriptive statistics, where the above described transformations are executed prior to calculating these statistics. Prior to estimating our models, we standardise all explanatory variables with their original in-sample mean and in-sample standard deviation, we do this once and do not update it during our rolling window estimation approach. We use standardised data to improve the estimation procedure for shrinkage methods. Additionally, we make use of in-sample data and out-of-sample data. Our in-sample data covers the period from February 3, 2000 till August 13, 2014 ($T_{IS} = 3653$), this sample is used to standardise our data and to perform the initial estimation of our models. Our out-of-sample data spans from August 14, 2014 till December 31, 2021 ($T_{OOS} = 1859$), and this sample is used to compute and evaluate forecasts of volatility.

Table 1: Descriptive statistics

	Mean	Median	Maximum	Minimum	Std. Deviation	Skewness	Kurtosis
RVD	2.52	1.24	166.56	0.12	5.16	13.52	374.28
RVW	2.52	1.30	76.56	0.20	4.34	7.66	96.07
RVM	2.53	1.35	40.90	0.30	3.73	5.20	39.26
RV^+	1.28	0.62	76.88	0.04	2.70	12.09	275.87
RV^-	1.26	0.61	86.40	0.04	2.65	13.91	461.30
M1W	0.00	0.00	0.29	-0.29	0.04	-0.09	7.01
$VOL	0.00	-0.01	2.49	-2.00	0.35	0.29	2.94
VIX	19.93	17.74	82.69	9.14	8.78	2.22	7.80
EPU	109.21	87.49	807.66	3.32	81.78	2.48	9.67
US3M	0.00	0.00	1.34	-1.60	0.05	-3.02	284.59
HSI	0.00	0.00	0.02	0.00	0.00	16.27	373.53
ADS	-30.13	-14.57	913.93	-2648.76	201.45	-7.69	93.63

Notes: The descriptive statistics for US3M and $VOL are calculated after the variables were transformed, ADS is in percentages.

3 Methodology

This section describes the models we use in our research, together with the metrics we use to evaluate the forecasts. We use a rolling window estimation approach in our research, where we re-estimate the models on a daily basis in order to produce the most accurate forecasts. We start forecasting on August 14, 2014, which gives us a window size of 3653 observations. In total, 1859 forecasts are made for each model.

The aim in this research is to estimate the quadratic variation of stock returns. However, the quadratic variation is not directly observable. Therefore, we estimate the realised variance as a indication for the quadratic variation. The derivation of the realised variance based on the five-minute log returns on an asset, $r^2_{t-j\cdot\Delta}$, is given by

$$RV_t = \sum_{j=1}^{R} r^2_{t-(j-1)\cdot\Delta}, \tag{1}$$

where R is the number of intraday observations of $r^2_{t-j\cdot\Delta}$ at day t ($R = 78$), and $\Delta = 1/R$ (Kleen and Tetereva, 2022). RV_t is known to be a consistent estimator of the quadratic variation, it is however not robust to microstructure noise (Andersen et al., 2011).

3.1 HAR Models

The HAR model by Corsi (2009) is a widely used benchmark model in volatility forecasting literature, e.g. Corsi et al. (2012); Santos and Ziegelmann (2014). Therefore, we use this model as benchmark in our research. The HAR model relies solely on realised variance as it regresses the RV on lagged variables of the RV. The regressors in the basic HAR model are the daily, weekly and monthly lags of the realised variance. These variables are given by $RV_t^d = RV_t$, $RV_t^{(w)} = \frac{1}{4}\sum_{i=1}^{4} RV_{t-i}$ and $RV_t^{(m)} = \frac{1}{17}\sum_{i=5}^{21} RV_{t-i}$. We define the lag variables in a decorrelated way to ensure minimal correlation between the variables. This leads to the specification of the HAR model as,

$$RV_t = \beta_0 + \beta_1 RV_{t-1}^{(d)} + \beta_2 RV_{t-1}^{(w)} + \beta_3 RV_{t-1}^{(m)} + \varepsilon_t, \tag{2}$$

where β_0 is the constant term of the model, β_i is the estimated coefficient of the explanatory variable i and ε_t the error term for observation t.

In extension to the basic HAR model, four additional HAR based models are used which are mostly in line with the models from Christensen et al. (2021). The first additional model is the log-version of the HAR model (logHAR), proposed by Corsi (2009). This model is used to better incorporate the possible nonlinear relationship between current and upcoming volatility. The model is defined as:

$$\log(RV_t) = \beta_0 + \beta_1 \log(RV_{t-1}^{(d)}) + \beta_2 \log(RV_{t-1}^{(w)}) + \beta_3 \log(RV_{t-1}^{(m)}) + \varepsilon_t. \tag{3}$$

The forecasts from this model are different from the regular HAR model as we need to take Jensen's inequality into consideration. The forecasts from this model are denoted by

$$\widehat{RV_{t+1}} = \exp(\hat{\beta}_0 + \hat{\beta}_1 \log(RV_{t-1}^{(d)}) + \hat{\beta}_2 \log(RV_{t-1}^{(w)}) + \hat{\beta}_3 \log(RV_{t-1}^{(m)}) + \frac{1}{2}\mathbb{V}(\varepsilon_t)) \tag{4}$$

Next, to accommodate for different effects of past positive and negative returns we include the semivariance HAR (SHAR) model in our analysis, introduced by Patton and Sheppard (2015). The model is denoted as:

$$RV_t = \beta_0 + \beta_1^- RV_{t-1}^- + \beta_1^+ RV_{t-1}^+ + \beta_2 RV_{t-1}^{(w)} + \beta_3 RV_{t-1}^{(m)} + \varepsilon_t, \tag{5}$$

where RV_{t-1}^- and RV_{t-1}^+ are defined as described in Barndorff-Nielsen et al. (2008).

As our third extension to the HAR model is the HARQ model by Bollerslev et al. (2016). This model takes into account the problem that arises due to the fact that the realised variance is a generated regressor. The model is specified as:

$$RV_t = \beta_0 + (\beta_1 + \beta_{1Q} RQ_{t-1}^{1/2}) RV_{t-1}^{(d)} + \beta_2 RV_{t-1}^{(w)} + \beta_3 RV_{t-1}^{(m)} + \varepsilon_t, \tag{6}$$

where $RQ_t = \frac{R}{3}\sum_{j=1}^{R} r_{t-(j-1)\cdot\Delta}^4$, is the realised quarticity of a stock.

In general, HAR models can include variables that are unrelated to past returns. In order to more fully compare our HAR models with our ML models, we consider the case where the HAR models use our full dataset. So, in addition to the variables mentioned above, each HAR model, except the basic HAR model, also includes all available predictors as described in Table 1. For example, the following vector represents the included variables for estimating the HARQ model $(RV_{t-1}^{(d)}, RQ_{t-1}^{1/2} RV_{t-1}^{(d)}, RV_{t-1}^{(w)}, RV_{t-1}^{(m)}, \text{M1W}_{t-1}, \ldots, \text{ADS}_{t-1})$, where the dots indicate all the variables included in Table 1. Moreover, we include an extended version of the basic HAR model which utilizes all predictors, denoted by HAR-X. This leaves us with five HAR models. Finally, the coefficients in the HAR models are estimated using a least squares approach.

3.2 Shrinkage Methods

The next set of models we investigate in this research are so called shrinkage methods or penalized regressions. Penalized regressions are generally formulated as

$$G_h(\mathbf{x}_t) = \beta_h' \mathbf{x}_t, \quad \widehat{\beta}_h = \arg\min_{\beta_h} \left[\sum_{t=1}^{T-h} \left(y_{t+h} - \beta_h' \mathbf{x}_t\right)^2 + \sum_{i=1}^{n} p\left(\beta_{h,i}; \lambda, \omega_i\right) \right], \tag{7}$$

where the penalty function is given by $p(\beta_{h,i}; \lambda, \omega_i)$; λ is the parameter of the penalty term; ω_i are weights on the penalty terms, it holds that $\omega_i > 0$.

In this paper we look at three different shrinkage methods with two corresponding adaptive versions of these methods. The difference between these methods lies within the penalty term. The general formula for the penalty is given as

$$\sum_{i=1}^{n} p\left(\beta_{h,i}; \lambda, \omega_i\right) := \alpha\lambda \sum_{i=1}^{n} \beta_{h,i}^2 + (1-\alpha)\lambda \sum_{i=1}^{n} \omega_i \left|\beta_{h,i}\right|. \tag{8}$$

The different models can be obtained by setting different values for α.

The first model we use is the Ridge Regression (RR) of Hoerl and Kennard (1970), this model is obtained by choosing $\alpha = 1$. Next we use the Least Absolute Shrinkage and Selection Operator (LA) from Tibshirani (1996). This method is achieved by setting $\alpha = 0$ and $\omega_i = 1$. The main difference between RR and LA is that LA can shrink the coefficients of variables to zero, whereas in RR the coefficients of the least important variables never become exactly zero. In addition, Zou (2006) proposes Adaptive LA (A-LA), which adaptively changes the weights, ω_i, of the penalty term based on the parameters of a first step LASSO estimation. Moreover, by setting $\omega_i = 1$ and $\alpha \in [0, 1]$ we obtain a convex combination of the two previous methods, this combination is called the Elastic Net (EN), proposed by Zou and Hastie (2005). In our research we use $\alpha = 0.5$ instead of an optimal partitioning due to feasibility constraints. Again, we use an adaptive version of the EN method, the adaptive EN (A-EN). This is obtained similarly to the A-LA method by adjusting the weights based on a first step estimation.

3.3 Tree-Based Regressions

The random forest model, proposed by Breiman (2001), comprises an ensemble of regression trees. In this algorithm, each tree is trained on a bootstrapped sample. We employ block bootstrapping, as this is used in the setting of time series forecasting. The algorithm employs B provided samples to estimate a tree with K_b regions. This is done based on a subset of the original features for each sample b, where $b = 1, .., B$, the subsets get randomly selected (Masini et al., 2023).

By repeating this step B times, we get b different models and take all models as a forecast combination. Several studies have already shown the efficiency of random forests in the field of volatility forecasting (Guo et al., 2018; Luong and Dokuchaev, 2018). With the help of bagging, the RF not only reduces the variance when averaging the regression trees but also reduces overfitting and improves the generalization of the model.

The forecasts of the RF algorithm are averages of the combination in each tree, denoted by

$$\widehat{RV}_{t+1} = \frac{1}{B} \sum_{b=1}^{B} \left[\sum_{k=1}^{K_b} \widehat{c}_{k,b} \boldsymbol{I}_{k,b} \left(\boldsymbol{x}_t; \widehat{\boldsymbol{\theta}}_{k,b} \right) \right], \tag{9}$$

where $\widehat{c}_k$ is the average of values that fall within a region R_k, $\widehat{\boldsymbol{\theta}}_{k,b}$ is the set of parameters that define the k-th region and $\boldsymbol{I}_{k,b}(\boldsymbol{x}_t; \widehat{\boldsymbol{\theta}}_{k,b})$ is an indicator function to determine if a set of predictors, given by $\boldsymbol{x}_t$, falls within a region R_k in the b-th sample (Medeiros et al., 2021).

The second tree-based model we include in our research is the extreme gradient boosting (XGBoost) model. Which is one of the most popular ML methods based on the principle of boosting. XGBoost is an extension of the gradient boosting technique by Friedman (2001), where the gradient boosting is performed by iteratively correcting the mistakes of the previous model to obtain a more precise model. However, XGBoost distinguishes itself from traditional gradient boosting by incorporating a regularization term and employing a second-order approximation of the loss function. According to Chen and Guestrin (2016), this reduces computational costs and helps to stop overfitting. The forecast is given by

$$\widehat{RV}_{t+1} = \sum_{k=1}^{t+1} f_k \left(x_t\right), \quad f_k \in F, \tag{10}$$

where F represents the space of the regression trees, and f_k denotes a single tree structure in F, with $t + 1$ additional functions to obtain a forecast as outcome.

3.4 Forecast Combinations

Forecast combinations of different models are widely studied in the academic literature, with the main question being which weights to use as optimal distribution (Timmermann, 2006; Smith and Wallis, 2009). In our research we employ an equal weighted distribution for each model, giving the forecast of model i an weight of $1/m$, with m being the number of models used in the combination. We use three different combinations of models in this research. The first is a combination of the five HAR models, denoted by FC$_{HAR}$. Second, we make a combination of the forecasts from the seven ML models, denoted by FC$_{ML}$. Third, we estimate a combination of forecasts from all the models described in this study, denoted by FC$_{all}$. The forecasts will then be given by

$$\widehat{RV}_{t+1} = \frac{1}{m} \sum_{i=1}^{m} \widehat{RV}_{i,t+1}, \tag{11}$$

where $\widehat{RV}_{i,t+1}$ is the forecast of the realised variance of model i at day $t + 1$.

3.5 Forecast Evaluation Methods

In our research, we use the Mean Squared Error (MSE) as a metric to evaluate forecasts. This is a widely used metric in the volatility forecasting literature, as it is one of the only two metrics for volatility forecasting that is robust to noise (Patton, 2011). The MSE for stock i and model j is defined as

$$\text{MSE}_{i,j} = \frac{1}{|OOS|} \sum_{t \in OOS} (RV_{t+1} - \widehat{RV}_{t+1})^2 \tag{12}$$

where $\widehat{RV}_{t+1}$ is the forecasted value of the realised variance and OOS denotes our out-of-sample period. However, for each model we obtain 24 MSEs, one for each one of the 24 stocks. To compare the models we thus need to compute the cross-sectional average relative MSE for each model, the cross-sectional average MSE of model s relative to model t is given by

$$\text{relMSE}_{s,t} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{MSE}_{i,s}}{\text{MSE}_{i,t}} \tag{13}$$

where N is the number of stocks. In addition, to check whether the MSE of two models is significantly different for the same stock, we use a pairwise Diebold-Mariano test between all models.

Moreover, we utilize SHAP values to interpret our ML models, introduced by Lundberg and Lee (2017). SHAP values are widely used as ML interpretation method across the existing academic literature. SHAP is derived from the Shapley value, which determines at what level each explanatory variable contributes to generating forecasts of a model. Furthermore, the Shapley values take interactions between variables into account (Molnar, 2020). We can use the SHAP framework for our tree-based models by utilizing the tree SHAP approach of Lundberg et al. (2018).

For analyzing the SHAP values we follow Joseph (2019) in using the Shapley regression framework, given by

$$RV_t = \Phi^S(x_i)\widehat{\beta}^S + \widehat{\varepsilon}_t, \quad \widehat{\varepsilon}_t \sim \mathcal{N}(0, \sigma_\varepsilon^2), \tag{14}$$

where $\widehat{\beta}_k^S$ are the surrogate model coefficients, for $k > 0$, which are evaluated against the null hypothesis $\mathcal{H}_0^k(\Omega) : \{\beta_k^S \leq 0|\Omega\}$. By using this framework we are able to evaluate the relation between the SHAP components of the variables and the realised variance. Rejecting the null hypothesis implies that the SHAP components supply us with statistically significant important information about the realised variance.

Finally, we use the Model Confidence Set (MCS) of Hansen et al. (2011) to provide a comprehensive comparison of all models. The MCS is a subset of all models that includes the superior models for which there is no statistical evidence to distinguish their forecasting performance. We follow Christensen et al. (2021) in applying the MCS with a confidence level of 0.9, which means that the best performing model is included in the MCS with 90% confidence. The number of bootstrapped samples is set to 5000 and the block length is equal to 22.

4 Results

4.1 Out-Of-Sample MSE Analysis

Table 2 shows the pairwise relative MSE of all our models, where the MSEs are computed as an relative average over all 24 stocks as described in Equation 13. The table can be read as follows. Reading by column, the first row represents the relative MSE of the averaged forecasted realised variance over all stocks relative to the standard HAR model, the second row represents the MSEs relative to the HAR-X model, and so on. Moreover, we compute the Diebold-Mariano statistic to test for significant differences in the MSEs. The number formatting in Table 2 shows the conclusion that can be drawn from the test, **number** (***number***) [***number***] shows whether the null of the Diebold-Mariano test is rejected for more than half of the stocks for a 10% (5%) [1%] confidence level. Where the null is rejected if the model in the column generates significantly better forecasts than the model in the row. Additionally, most of the results are being compared to the MSE of the HAR model, the 24 MSEs for the HAR model can be found in Appendix D. In addition, we supply the MSEs of the RR, RF and combined forecasts in the same table.

Interesting results emerge from Table 2. First, we observe that the HARQ model is the best performing HAR model, with the HARQ model reducing the MSE by 3.6% relative to the HAR. The HARQ model makes use of a company's realised quarticity (RQ), which measures the cumulative deviation of volatility over a given period (Mancino and Sanfelici, 2012). The RQ is used to gain insight into the persistence of volatility for a time series. In addition, the HARQ model allows the parameters to vary with the degree of measurement error. In this way, the model provides more accurate forecasts for periods with low measurement error (Bollerslev et al., 2016).

Table 2: Relative MSEs and Diebold-Mariano statistic of the one-day-ahead forecasts from our models

	HAR	HAR-X	logHAR	SHAR	HARQ	RR	LA	EN	A-LA	A-EN	RF	XGB	FC$_{HAR}$	FC$_{ML}$	FC$_{all}$
HAR	-	0.986	1.852	1.031	0.964	0.963	0.986	0.986	0.985	0.985	0.971	1.278	0.922	0.926	0.897
HAR-X	1.269	-	1.892	1.046	0.984	0.978	0.998	0.998	1.000	1.001	0.980	1.254	***0.943***	0.937	***0.913***
logHAR	0.772	0.749	-	0.785	0.743	**0.736**	0.753	0.753	0.756	0.756	**0.730**	0.927	**0.703**	**0.707**	**0.687**
SHAR	0.994	0.968	1.825	-	0.952	0.943	0.963	0.963	0.965	0.965	0.942	1.204	**0.910**	0.903	**0.881**
HARQ	1.049	1.027	1.949	1.073	-	1.004	1.027	1.026	1.027	1.028	1.010	1.316	0.961	0.962	0.934
RR	1.055	1.028	1.956	1.071	1.012	-	1.021	1.021	1.022	1.022	1.000	1.280	0.969	0.958	**0.935**
LA	1.037	1.008	1.922	1.051	0.993	0.981	-	1.000	1.001	1.001	0.982	1.248	0.952	**0.939**	**0.918**
EN	1.037	1.009	1.922	1.051	0.993	0.981	1.000	-	1.001	1.002	0.982	1.247	0.952	**0.939**	**0.918**
A-LA	1.035	1.009	1.923	1.051	0.993	0.981	1.000	1.000	-	1.001	0.984	1.253	0.952	**0.939**	**0.918**
A-EN	1.035	1.008	1.924	1.050	0.992	0.980	0.999	0.999	0.999	-	0.983	1.252	0.951	**0.938**	**0.918**
RF	1.099	1.065	2.049	1.107	1.052	1.035	1.057	1.056	1.058	1.058	-	1.253	1.005	0.984	0.964
XGB	0.955	0.914	1.787	0.948	0.907	0.887	0.901	0.900	0.903	0.904	0.842	-	0.869	0.836	0.827
FC$_{HAR}$	1.092	1.071	1.950	1.117	1.047	1.047	1.072	1.072	1.072	1.073	1.050	1.374	-	1.004	0.973
FC$_{ML}$	1.110	1.078	2.054	1.122	1.061	1.048	1.069	1.069	1.070	1.071	1.040	1.317	1.016	-	0.978
FC$_{all}$	1.130	1.102	2.085	1.148	1.081	1.074	1.098	1.097	1.098	1.099	1.070	1.378	1.034	1.027	-

Note: We report the MSE of the forecasted values for the realised variance. The MSE is computed as the relative average across all 24 stocks as described in Section 3.5. The formatting for the Diebold-Mariano test is done as follows: **number** (***number***) [***number***] denotes whether the null hypothesis of equal forecasting accuracy is rejected more than 50% of the time at a 10% (5%) [1%] significance level. The hypothesis being tested is that the out-of-sample MSE of the model in a selected row is significantly higher than the MSE of the model in a selected column, where the out-of-sample period starts on August 14, 2014 and ends on December 31, 2021.

Moreover, the only other HAR model outperforming the standard HAR is the HAR-X model, which reduces the MSE by 1.4% relative to the HAR. Looking at our remaining two HAR models, logHAR and SHAR, we find that these models perform worse than the basic HAR model. In addition, the logarithmic transformation of the realised variance provides a more symmetric distribution for the model's error terms, as they become closer to a Gaussian distribution. In our research, we find that the logHAR model struggles to respond quickly to sudden shocks in the realised variance. This is, together with the fact that the Covid-19 pandemic is included in our out-of-sample period, where many rapid changes in volatility occurred, the possible reason for the poor performance of the logHAR in our research.

Furthermore, Christensen et al. (2021) argues that the poor performance of the SHAR model may be due to overfitting, as they claim that lack of regularisation makes HAR models susceptible to in-sample overfitting when additional variables are employed. However, as mentioned above, the HAR-X model does outperform the HAR model, which implies that adding more features to the regular HAR model does improve its forecasting accuracy. Therefore, the poorer performance of the SHAR is more likely to be due to the predictive incompetence of the positive and negative semivariances of the realised variance. Furthermore, the SHAR model presumably fails to capture the leverage effect in our case. Which refers to the negative correlation between the return on a stock and the volatility tendency (Ait-Sahalia et al., 2013).

The shrinkage models are expected to alleviate the potential problem of overfitting, as this set of models penalises high values of the estimated coefficients. This is confirmed by Table 2, where we see that RR reduces the MSE by 3.7% relative to the regular HAR model. However, RR does not provide a MSE reduction relative to the HARQ model. In addition, LA and EN reduce the MSE by 1.4% and thus fare worse than RR. Their adaptive versions perform slightly better, but still worse than RR. Furthermore, as the difference in MSE reduction between the adaptive LA and EN models and the regular LA and EN models is negligible, we conclude that adaptively adjusting the strength of the penalty term of these models does not lead to more accurate forecasts in our case. In conclusion, the best performing linear model is the RR. However, the RR does not significantly outperform the other linear models, including the HAR model, as indicated by the Diebold-Mariano test.

Next we examine the results of the tree-based models. Looking at the results of the XGB model, we see that it is the worst performing model next to the logHAR model, with a relative increase in MSE of 27.8% compared to the standard HAR model. We see that the XGB model fails to counteract overfitting, although the second order approximation of the loss function is used to prevent this. The cause of this disappointing performance may lie in the regular gradient boosting algorithm, which allocates more weight to outlying observations. Given that our out-of-sample period includes the Covid-19 pandemic, where many outliers occurred, and the aforementioned property of the gradient boosting algorithm, it is not entirely surprising that the XGB model does not provide us with accurate forecasts of volatility. Our second tree based model, the RF model, performs contradictory to the XGB model, as the RF model outperforms all HAR models besides the HARQ model. The RF model reduces the MSE by 2.9% compared to the HAR model. The accurate performance of RF is due to its strength in decorrelating trees by randomly selecting the set of explanatory variables. It should be noted that the difference in the reduction of the MSE between the RF and all models, except the logHAR model, is not statistically significant for more than 50% of the stocks.

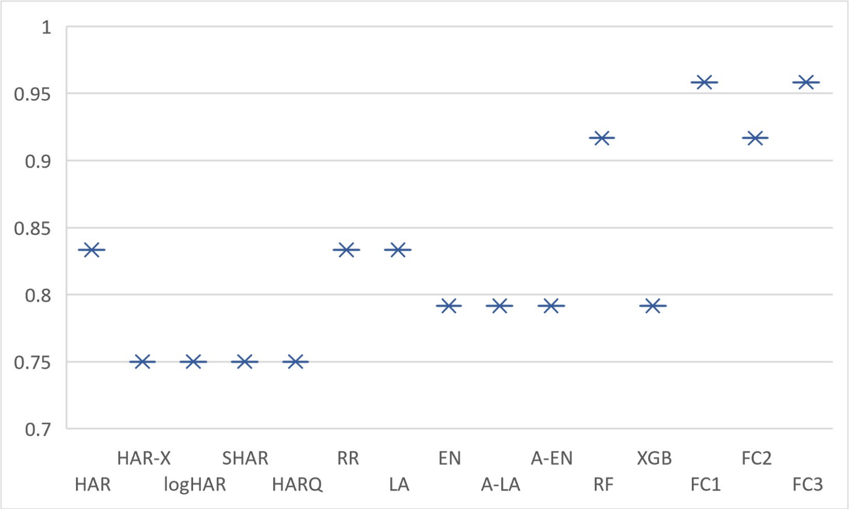
Finally, we evaluate the combined forecasts produced by our previous models. First, we look at the combination of the HAR models, FC$_{HAR}$. Table 2 shows that this combination method yields a reduction in MSE of 7.8%, outperforming all models evaluated so far. Compared to the extensions of the HAR model, we see a significant improvement in terms of MSE for more than 50% of the stocks for three out of four extensions. However, this combination does not significantly outperform the regular HAR model. Comparing this model with the shrinkage models again results in more accurate forecasts according to the MSE. Here, the combination of HAR models yields a MSE reduction of around 3% to 5% compared to the shrinkage models.

Next, by looking at the MSE of the forecast combination including only ML models, FC$_{ML}$, we conclude that this combination performs slightly worse than FC$_{HAR}$. However, FC$_{ML}$ still reduces the MSE by 7.4% compared to the HAR. The remaining results for this model are relatively similar to those of FC$_{HAR}$. However, unlike the combination of HAR models, this model significantly outperforms four out of five shrinkage models for more than 50% of the stocks examined at a 10% significance level. The RR model is the only shrinkage method not outperformed by FC$_{ML}$. Finally, the combination of all models, FC$_{all}$, is the best performing of our three combinational models, reducing the MSE by 10.3%, outperforming all other models. Additionally, it significantly outperforms all shrinkage models, even RR, and three HAR models, again the regular HAR model is not significantly outperformed. The predictive accuracy of the combined forecasts can be inferred from the properties of forecast combination methods. As these methods combine different models to improve accuracy, reduce uncertainty and compensate for individual model errors, resulting in more robust and reliable predictions (Timmermann, 2006).

From the results based on the MSEs we conclude that the best performing method to obtain accurate forecasts of the realised variance is combining forecasts. Therefore, we recommend investors, traders and market participants to use this method for predicting volatility on stocks.

Next, Figure 1 shows the ratio at which each model is included in the MCS for a 90% confidence set.

Figure 1: MCS inclusion rates for each model, computed using a 90% confidence set



Notes: We report the MCS inclusion rate for each model. The number of bootstraps is set to 5000, the block bootstrap length is set to 22 and the confidence level is equal to 90%. Indicating that the best performing model is included in the MCS with 90% confidence.

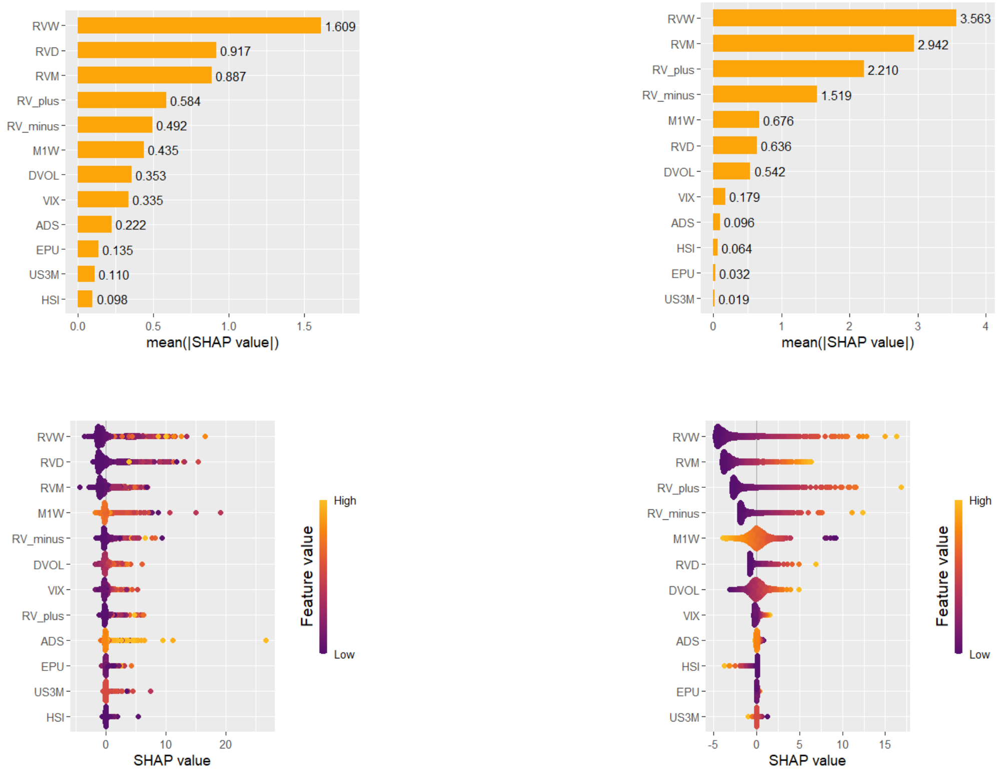
From the figure we observe the following. First, we see that the inclusion rate for all shrinkage models varies between 79% and 84%, being similar to the inclusion rate of the HAR model. Furthermore, the RF has an inclusion rate of 92%, which is, besides the FC models, the highest rate. Finally, the rates of the FC models are, as expected, very high, with the rates of the FC$_{HAR}$ and FC$_{all}$ model being equal to 96%, corresponding with the results obtained from our MSE analysis.

4.2 Machine Learning Interpretability

In this section we analyse the RR and RF models using SHAP values. We analyse these models as from Section 4.1 we conclude that these are the two best performing ML models based on the MSE. Moreover, we need to address two important notes regarding these figures. Firstly, the SHAP values are calculated for a single stock, we have arbitrarily chosen Apple (AAPL) in our case. Therefore, this figure is not representative of the entire sample. For robustness, we reproduce the same figure for the Nike stock in Appendix E. Secondly, these plots do not necessarily imply a causal relation between the explanatory and dependent variables, as they solely denote an indication of the behaviour of the variables given our data sample.

The top two graphs in Figure 2 show that lagged realised variances are the most important

Figure 2: Feature importance and SHAP summary plots



(a) Random Forest

(b) Ridge Regression

Notes: The top two sub-figures show the mean SHAP value for all explanatory variables, computed for the Apple Inc. stock. In the bottom two sub-figures, each point represents an observation with its corresponding SHAP value on the x-axis, these are the so-called SHAP summary plots. The colour scale on the sides represents the actual values of the characteristics.

features for both RR and RF, especially the weekly lag. This result can be used to argue in favor of the effectiveness of HAR models, and can also explain why our ML models do not show large reductions in MSE relative to the HAR models. Additionally, by looking at Table 4 in Appendix C we observe that for the HAR-X model the most important variables correspond with the results for RR and RF. As for the HAR-X model, the coefficients for the lagged realised variances obtain the lowest p-values when performing an OLS regression. This affirms that there exists a relation between past and future volatility. Furthermore, the lagged realised variances are followed by the remaining two firm-specific variables, M1W and \$VOL, when looking at the highest p-values. This again highlights the similar functioning of the HAR and ML models.

Furthermore, forecasts by RR are relatively more affected by the lagged features than RF forecasts. This is evident from the mean SHAP values, which are consistently about two to three times higher for RR than for RF. It is noticeable that the daily lagged realised variance is relatively less important for RR than for RF, otherwise the ranking is quite similar. Furthermore, we see that the firm-specific variables all have a higher average SHAP value than the macroeconomic variables. This implies that the latter play a significantly smaller role in forecasting realised variance, which again argues in favor of the HAR model. Among the firm-specific variables, one-week momentum (M1W) and dollar trading volume ($VOL) play the smallest role in predicting realised variance. Finally, among the macroeconomic variables the VIX is the most important predictor for realised variance. The ADS, HSI, EPU and US3M seem to play a negligible role in forecasting volatility.

The remaining two plots provide a better understanding of the contribution of the variables to the forecasts. For example, a low value for any of the lagged realised variances has a negative impact on the SHAP values of both models. This implies that such observations contribute little to the forecast values of the realised variance. Furthermore, we see that the M1W shows an opposite pattern, with low values of the characteristic having a large contribution to the forecasts and high values having a small contribution. For RF, the contribution of the ADS feature does not seem to depend on the value of the feature, as the orange dots are spread over the entire x-axis. The rest of the variables seem to behave similarly to the lagged realised variances, with small negative contributions at low values and larger positive contributions at higher values.

Given the behaviour of the SHAP values for M1W, we would expect the corresponding estimated coefficient of this characteristic for the HAR-X model to be significantly negative. Looking at Table 4 we see that this is indeed the case, as M1W has the only significantly negative coefficient. Comparing the other coefficients with the bottom two plots in Figure 2, we see similar results. For example, the coefficients of the lagged realised variances are all significantly positive, as would be expected from our conclusion above. In addition, the coefficient on the ADS index is very close to zero, with a p-value of only 0.016, consistent with the negligible effect of this feature on the SHAP values.

Finally, for robustness, we compare the results obtained from the SHAP values as described in Figure 2 with the results for the Nike stock in Appendix E, we see that the results are largely consistent. The VIX however seems to have a relatively higher impact for this stock, suggesting that it can be useful to include macroeconomic variables.

5 Conclusion

This paper evaluates the forecasting ability of several volatility forecasting models. This is examined by evaluating five HAR models as benchmarks and comparing their forecasts with five shrinkage models, two tree-based machine learning approaches, and three sets of combined forecasts from other models. We show that the combined forecasts provide the most accurate predictions of realised variance, which is concluded by using the MSE as evaluation criterion. The combined forecasts from all models generate the highest reduction in MSE, 10.3%. The accurate forecasting ability of the combined forecasts is due to their ability of reducing uncertainty and compensating for individual model errors. Additionally, we look at the inclusion rate in the MCS, from which we observe quite consistent results, as the combined forecasts are included in the MCS with the highest rate.

In addition to evaluating the forecasting ability of the models, we contribute by adding interpretation to two of our ML models. We use SHAP for this purpose. In particular, we utilise this method to interpret the forecasts made by the RF and RR models, as these are our best performing ML models. An examination of the SHAP values indicates that the lagged values of the realised variance are the most influential features in forecasting stock volatility. This suggests that the ML methods are based on the same fundamentals as the HAR models. Furthermore, we see that the lagged values of the realised variance are followed by the other firm-specific variables as the most influential features, while the macroeconomic variables are the least important. This suggests that stock volatility mainly depends on the characteristics of the respective firm. Whereas, macroeconomic variables play a less important role in forecasting the realised variance.

We acknowledge some limitations to our research. Firstly, Christensen et al. (2021) found that neural network approaches outperformed RF for one-day-ahead volatility forecasting. Therefore, further research could consider neural networks as part of their analysis. Furthermore, in our research we use SHAP to investigate the interpretation behind ML methods, there are however several other ML interpretability methods that can be used to investigate the 'black box' surrounding ML approaches. Some examples are Accumulated Local Effects by Apley and Zhu (2020) and the Model Class Reliance approach by Fisher et al. (2019). Moreover, this research does not extensively tune its hyperparameters, for further research it would be interesting to observe whether thorough hyperparameter tuning leads to significant differences in results. Finally, this research only employs the MSE to evaluate the employed models. For robustness, the QLIKE statistic could be additionally employed in future research, as Patton (2011) argues that the QLIKE is robust to positive outliers and heavily influenced by negative outliers. Contradictory, the MSE is robust to negative outliers and strongly influenced by positive outliers.

References

Y. Ait-Sahalia, J. Fan, and Y. Li. The leverage effect puzzle: Disentangling sources of bias at high frequency. *Journal of Financial Economics*, 109(1):224–249, 2013.

T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys. Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625, 2003.

T. G. Andersen, T. Bollerslev, and N. Meddahi. Realized volatility forecasting and market microstructure noise. *Journal of Econometrics*, 160(1):220–234, 2011.

D. W. Apley and J. Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82 (4):1059–1086, 2020.

S. B. Aruoba, F. X. Diebold, and C. Scotti. Real-time measurement of business conditions. *Journal of Business & Economic Statistics*, 27(4):417–427, 2009.

F. Audrino and S. D. Knaus. Lassoing the har model: A model selection perspective on realized volatility dynamics. *Econometric Reviews*, 35(8-10):1485–1521, 2016.

S. R. Baker, N. Bloom, and S. J. Davis. Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4):1593–1636, 2016.

O. E. Barndorff-Nielsen, S. Kinnebrock, and N. Shephard. Measuring downside risk-realised semivariance. *CREATES Research Paper*, 1(42), 2008.

G. Bekaert and M. Hoerova. The vix, the variance premium and stock market volatility. *Journal of econometrics*, 183(2):181–192, 2014.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986.

T. Bollerslev, A. J. Patton, and R. Quaedvlieg. Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, 192(1):1–18, 2016.

L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

A. Bucci. Realized volatility forecasting with neural networks. *Journal of Financial Econometrics*, 18(3):502–531, 2020.

M. Caporin and F. Poli. Building news measures from textual data and an application to volatility forecasting. *Econometrics*, 5(3):35, 2017.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

K. Christensen, M. Siggaard, and B. Veliyev. A machine learning approach to volatility forecasting. *Available at SSRN*, 2021.

P. F. Christoffersen and F. X. Diebold. How relevant is volatility forecasting for financial risk management? *Review of Economics and Statistics*, 82(1):12–22, 2000.

F. Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009.

F. Corsi and R. Renò. Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling. *Journal of Business & Economic Statistics*, 30 (3):368–380, 2012.

F. Corsi, F. Audrino, and R. Renó. HAR modeling for realized volatility forecasting. In *Handbook of volatiltiy models and their applications*, pages 363–382. John Wiley & Sons, Inc, 2012.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50:987–1007, 1982.

R. F. Engle and A. J. Patton. What good is a volatility model? *Quantitative Finance*, 1(2): 237–245, 2001.

A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

T. Guo, A. Bifet, and N. Antulov-Fantulin. Bitcoin volatility forecasting with a glimpse into buy and sell orders. In *2018 IEEE international conference on data mining (ICDM)*, pages 989–994. IEEE, 2018.

P. R. Hansen and A. Lunde. A forecast comparison of volatility models: does anything beat a garch (1, 1)? *Journal of applied econometrics*, 20(7):873–889, 2005.

P. R. Hansen, A. Lunde, and J. M. Nason. The model confidence set. *Econometrica*, 79(2): 453–497, 2011.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

A. Joseph. Parametric inference with universal function approximators. *[Online]. Available: http://arxiv.org/abs/1903.04209*, 2019.

L. Khaidem, S. Saha, and S. R. Dey. Predicting the direction of stock market prices using random forest. *[Online]. Available: http://arxiv.org/abs/1605.00003*, 2016.

O. Kleen and A. Tetereva. A forest full of risk forecasts for managing volatility. Technical report, Working Paper, 2022.

B. Lei, Z. Liu, and Y. Song. On stock volatility forecasting based on text mining and deep learning under high-frequency data. *Journal of Forecasting*, 40(8):1596–1610, 2021.

X. Li, C. Liang, and F. Ma. Forecasting stock market volatility with a large number of predictors: New evidence from the MS-MIDAS-LASSO model. *Annals of Operations Research*, 311(2): 1–40, 2022.

B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.

S. M. Lundberg and S. I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

S. M. Lundberg, G. G. Erion, and S. I. Lee. Consistent individualized feature attribution for tree ensembles. *[Online]. Available: http://arxiv.org/abs/1802.03888*, 2018.

C. Luong and N. Dokuchaev. Forecasting of realised volatility with the random forests algorithm. *Journal of Risk and Financial Management*, 11(4):61, 2018.

M. E. Mancino and S. Sanfelici. Estimation of quarticity with high-frequency data. *Quantitative finance*, 12(4):607–622, 2012.

R. P. Masini, M. C. Medeiros, and E. F. Mendes. Machine learning advances for time series forecasting. *Journal of economic surveys*, 37(1):76–111, 2023.

M. C. Medeiros, G. F. Vasconcelos, Á. Veiga, and E. Zilberman. Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39(1):98–119, 2021.

C. Molnar. *Interpretable machine learning.* Lulu Press, 2020.

P. Newbold and D. I. Harvey. Forecast combination and encompassing. In *A companion to economic forecasting.* Blackwell, Oxford, 2002.

A. J. Patton. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256, 2011.

A. J. Patton and K. Sheppard. Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics*, 97(3):683–697, 2015.

E. Rahimikia and S. H. Poon. Machine learning for realised volatility forecasting. *Alliance Manchester Business School, University of Manchester. Working paper*, 2020.

D. G. Santos and F. A. Ziegelmann. Volatility forecasting via midas, har and their combination: An empirical comparative study for ibovespa. *Journal of Forecasting*, 33(4):284–299, 2014.

J. Smith and K. F. Wallis. A simple explanation of the forecast combination puzzle. *Oxford bulletin of economics and statistics*, 71(3):331–355, 2009.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

A. Timmermann. Forecast combinations. *Handbook of Economic Forecasting*, 1:135–196, 2006.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Appendices

A Data source overview

Table 3: Data summary

	Acronym	Source	Transformation
Firm specific	RVD	Kleen and Tetereva (2022)	none
	RVW	Kleen and Tetereva (2022)	none
	RVM	Kleen and Tetereva (2022)	none
	RV^+	Kleen and Tetereva (2022)	none
	RV^-	Kleen and Tetereva (2022)	none
	\$VOL	Yahoo finance [1]	$\log(\frac{x_t}{x_{t-1}})$
	M1W	Yahoo Finance	none
Macroeconomic	EPU	FRED (USEPUINDXD)	none
	VIX	Kleen and Tetereva (2022)	none
	US3M	FRED (DTB3)	$x_t - x_{t-1}$
	HSI	Google Finance [2]	$(\log(\frac{x_t}{x_{t-1}}))^2$
	ADS	Federal Reserve Bank of Philadelphia	none

B Implementation Code and Tuning Parameters

All models are estimated in R using R packages. For model evaluation, MCS is performed in R using the `rugarch` package and the Diebold-Mariano test is performed using the `forecast` library. For our ML interpretability we use the `kernelshap` and `treeshap` packages from R. Furthermore, we use the `randomForest` package with all default tuning parameters, except from the number of trees which we set to 50 due to running time feasibility. For XGBoost we use the `xgboost` package in R, here we again use the default values, except nrounds = 1000, eta = 0.05, nthread = 1, colsample_bylevel = 2/3 , and max_depth = 4.

[1] E.g. https://finance.yahoo.com/quote/AAPL/history?p=AAPL

[2] https://www.google.com/finance/quote/HSI:INDEXHANGSENG?hl=nl

C Parameter estimates for the HAR(-X) model

Table 4: Parameter estimates of the Apple stock for the HAR(-X) models

	HAR	HAR-X
β_0	5.771 (62.976)	5.771 (66.156)
β_{RVD}	3.137 (7.397)	3.193 (6.926)
β_{RVW}	2.307 (4.815)	2.208 (4.792)
β_{RVM}	1.325 (3.920)	1.155 (3.202)
β_{M1W}	-	-0.852 (-5.215)
$\beta_{\text{\$VOL}}$	-	1.417 (12.019)
β_{VIX}	-	0.287 (1.055)
β_{US3M}	-	-0.178 (-1.651)
β_{EPU}	-	-0.075 (-0.460)
β_{HSI}	-	0.134 (0.517)
β_{ADS}	-	0.003 (0.016)

Notes: The HAR models are estimated for the in-sample period, starting on February 3, 2000 and ending on August 14, 2014. The coefficients are estimated using OLS, and their standard deviation is computed using Hubert-White heteroscedasticity-robust standard errors. The value in parentheses denotes the t-statistic of the coefficient.

D Raw MSEs

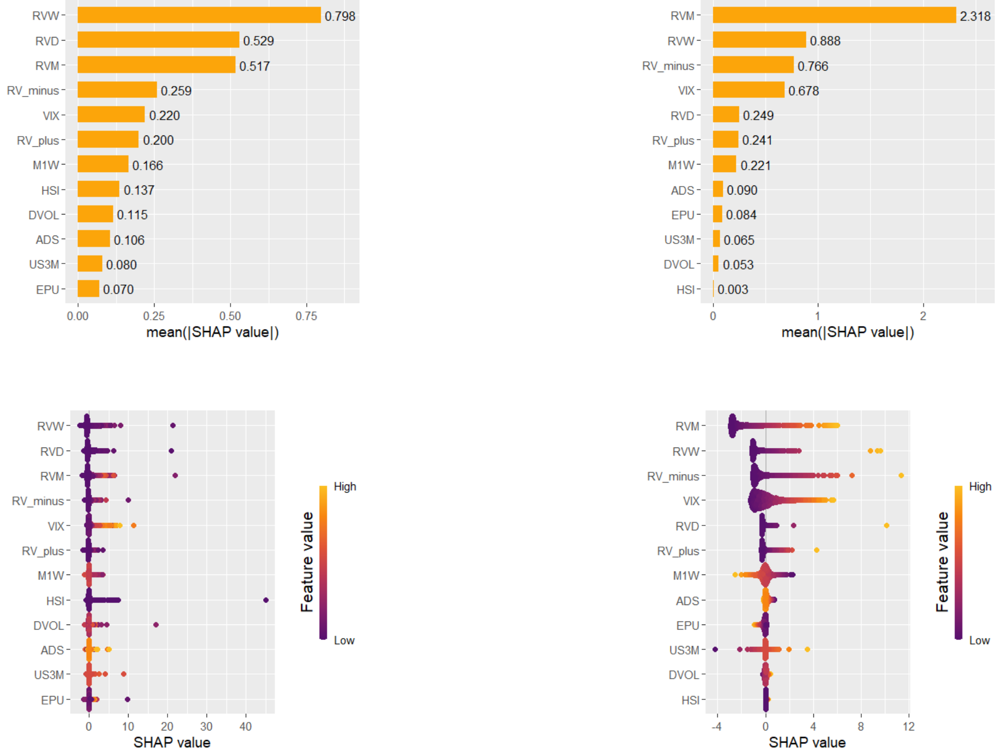
Table 5: MSEs for the RR, RF and combined forecasts

Company	HAR	RR	RF	FC1	FC2	FC3
MMM	3.34	3.23	3.82	2.41	3.11	2.66
AAPL	7.29	7.16	8.30	7.09	7.18	7.02
BA	82.31	77.84	62.44	76.98	75.76	75.86
CAT	5.15	4.80	4.74	4.92	4.78	4.78
CVX	7.67	7.64	6.57	6.83	7.19	6.83
CSCO	4.79	4.00	3.79	4.45	3.84	3.98
KO	3.47	2.86	2.92	3.47	2.79	2.98
DIS	5.35	5.10	4.35	4.72	4.57	4.57
XOM	5.39	5.34	6.57	4.96	5.39	5.02
GE	18.89	18.26	17.85	17.56	17.23	17.18
GS	5.23	7.61	7.70	5.57	7.88	6.34
HD	25.84	22.69	22.24	21.87	21.43	21.38
IBM	2.71	2.47	2.07	2.63	2.19	2.32
INTC	7.34	6.56	6.35	6.99	6.26	6.39
JNJ	5.23	6.16	4.79	5.14	5.46	5.21
MCD	5.76	4.31	3.17	4.42	3.41	3.70
MRK	4.57	4.43	5.13	4.33	4.43	4.36
MSFT	4.91	4.78	4.77	4.72	4.69	4.65
NKE	6.57	5.77	5.00	7.52	5.45	5.59
PFE	4.14	3.90	3.89	3.91	3.77	3.79
RTX	10.33	9.07	7.27	8.94	8.07	7.99
TRV	3.65	4.02	7.26	4.14	4.12	4.09
VZ	12.76	11.75	11.43	12.04	11.39	11.60
WMT	4.56	4.50	4.87	4.44	4.77	4.35

Notes: The values represent the MSE for each stock for six different models. These models are chosen as they show the most accurate results as described in Section 4. The HAR model is shown as this is our main benchmark model.

E SHAP robustness check

Figure 3: Feature importance and SHAP summary plots



(a) Random Forest

(b) Ridge Regression

Notes: The top two sub-figures show the mean SHAP value for all explanatory variables, computed for the Nike stock. In the bottom two sub-figures, each point represents an observation with its corresponding SHAP value on the x-axis, these are the so-called SHAP summary plots. The colour scale on the sides represents the actual values of the characteristics.