Handelshøyskolen BI

GRA 19703 Master Thesis

Thesis Master of Science 100% - W

Predefinert informasjon

Startdato:	09-01-2023 09:00 CET	**Termin:**	202310
Sluttdato:	03-07-2023 12:00 CEST	**Vurderingsform:**	Norsk 6-trinns skala (A-F)
Eksamensform:	T		
Flowkode:	202310||11184||IN00||W||T		
Intern sensor:	(Anonymisert)		

Deltaker

Navn:	Sander André Pilskog Stadsvik og Emil Andre Ås

Informasjon fra deltaker

Tittel *:	Forecasting Realized Volatility with Earnings Announcements and Overnight Returns
Navn på veileder *:	Paolo Giordani

Inneholder besvarelsen konfidensielt materiale?:	Nei	**Kan besvarelsen offentliggjøres?:**	Ja

Gruppe

Gruppenavn:	(Anonymisert)
Gruppenummer:	200
Andre medlemmer i gruppen:	

Forecasting Realized Volatility with Earnings Announcements and Overnight Returns

Master Thesis

by
Emil Ås and Sander Stadsvik
MSc in Quantitative Finance

Supervisor:
Paolo Giordani

Oslo, June 29, 2023

ABSTRACT

In our study, we forecast realized volatility utilizing a large panel of stocks from the S&P 500, with the inclusion of overnight returns and earnings announcements. Our comparative analysis employs both the heterogeneous autoregressive model and gradient boosting. Upon evaluation, we ascertain that the inclusion of earnings announcements moderately enhances the precision of RV forecasting. Furthermore, our findings suggest that the gradient-boosting methodology demonstrates superior performance in comparison to the HAR model.

Acknowledgements

The authors responsible for the creation of this thesis wish to extend our deepest gratitude to our esteemed supervisor, Paolo Giordani, who is a renowned figure in the Department of Finance at BI Norwegian Business School. His invaluable guidance and unwavering support were integral throughout our research journey.

Contents

List of Figures

List of Tables

1 Introduction and Motivation

"Prediction is very difficult, especially if it is about the future" (Niels Bohr).

In our research, we explore the potential of earnings announcements (EA) and overnight (ON) returns to enhance the prediction of realized volatility (RV). Given that earnings are announced in advance and typically released before the market opens or after it closes, we posit that they can increase the accuracy of forecasts, particularly in the periods immediately preceding and following the earnings announcement. To more effectively capture this effect, we propose that incorporating the squared ON log return will enhance the traditional academic definition of RV, resulting in forecasts of greater practical utility. Refer to Equation (29) for the RV measure used in our study. Our examination utilizes two cutting-edge models: the linear heterogeneous autoregressive (HAR) model, and the gradient boosting (GB) model, Light Gradient Boosting Machine (LightGBM).

We perform an in-sample analysis and compare the out-of-sample (OOS) performance of RV at different horizons on a large panel of individual stocks in the S&P 500. Our dataset includes intraday prices from both regular and extended hours, spanning January 3, 2005, to September 21, 2022. Compared to linear models, GB exhibits a superior ability to discern interaction effects and nonlinearities. Furthermore, we posit that the introduction of large-scale data will significantly enhance performance, with the impact being exponentially greater in a GB context compared to a panel regression setting. This hypothesis is supported by its demonstrated efficiency in handling big data, a claim substantiated by studies such as those by Bollerslev et al. (2018) and Li and Tang (2022). As a result, it has the potential to outperform the linear HAR model. We extend the analysis by conducting additional feature selection and engineering, based on stylized facts of RV and some empirical findings.

The volatility of financial markets and assets, usually defined as fluctuations in asset prices, holds major importance for financial market participants. This importance is evident for asset and risk managers, arbitrageurs, market makers, traders, insurers, and option pricing, among others. Quantitative asset managers use expected volatility for position sizing and other risk measures such as VaR and CVaR, which are crucial for position sizing to avoid potential leverage disasters. An example of such a disaster is Long-Term Capital Management (LTCM), a highly leveraged hedge fund led by Myron Scholes and Robert C. Merton (both Nobel laureates in Economics in 1997), which focused on convergence arbitrage. Despite their initial excellent performance, they underestimated the risk of wors-

ening mispricing, equatable to downside volatility risk. The 1997 Asian financial crisis and the onset of the 1998 Russian financial crisis led to a flight to liquidity, causing huge losses and a dramatic increase in their net leverage ratio. Ultimately, they could not meet their liabilities and were bailed out by a total of 14 banks for a total of $3.6 billion (Kabir & Hassan, 2005). For more information on this topic, the reader may refer to "When Genius Failed - The Rise and Fall of Long-Term Capital Management" (Lowenstein, 2001). Fundamental asset managers weigh risk against potential reward. Much of the finance literature is based on the notion of higher risk, higher reward (e.g., CAPM), although recent evidence challenges this theory. Accurate forecasting of expected volatility over the life of an option is essential for options arbitrage. The list of examples is extensive, underscoring the clear importance of volatility. Measures of volatility are crucial ex-post, but often more so ex-ante. Therefore, we need forward-looking estimates of RV, representing the actual volatility experienced by an asset over a given period.

The derivatives market, particularly options pricing, motivates our thesis. This vast and complex market involves the trading of financial instruments that derive their value from an underlying asset. Such instruments include futures contracts, options contracts, swaps, and more complex derivatives, like exotic options. The size of the derivatives market has seen significant growth in recent years; the Bank for International Settlements (BIS) estimated its notional size to be $632 trillion in 2022, with a gross market value of $18.3 trillion (Bank for International Settlements, 2022). This growth can be attributed to several factors: the need for risk management by financial institutions, the increasing demand for hedging strategies by investors, and the advent of new financial instruments that allow for more efficient risk transfer.

The history of volatility forecasting research is extensive, encompassing a variety of models. Initial forecasts of volatility often relied on implied volatility or assumed constant volatility, using either the sample average or a moving average of historic volatility. Generalized Autoregressive Conditional Heteroscedasticity (GARCH) models (Bollerslev, 1986) and stochastic volatility models were among the first rigorous methods for modeling RV, with GARCH models shown to effectively capture volatility clustering in the data. The HAR model (Corsi, 2009), a parsimonious linear model easily estimated by OLS, came later. Due to their accuracy and simplicity, HAR and its extensions are among the most widely used models for forecasting RV today. They successfully capture the persistence often seen in volatility data.

Options pricing relies on the RV experienced throughout the entire day. The standard academic definition, however, is calculated using data from regular trading hours and does not capture the volatility observed in extended market hours. We construct a slightly different measure of RV that we believe possesses a better capability to capture the true underlying daily distribution. This is achieved by incorporating the squared ON log return into the RV measure. To our knowledge, there is scant literature using this same procedure. Furthermore, we posit that another crucial aspect of volatility forecasting has been overlooked. Days of EA are well-known to cause fluctuations in stock prices. Earnings are generally announced during extended hours, so we believe our newly proposed measure of volatility will more effectively capture this effect. We include dummy variables for days of announcements, which are known to market participants months in advance, to assist the model in learning its impact.

Machine learning has seen tremendous growth over the past decade, resulting in the emergence of many promising models. Some authors have explored its use in forecasting volatility; however, most research has focused on neural networks, random forests, and regularization and variable selection techniques. Currently, there is limited literature on how GB performs in RV forecasting. Although the signal-to-noise ratio (SNR) in volatility data is high, we believe it is important to investigate further, as GB is known to perform well with most data, especially at capturing nonlinearities that HAR-type models do not adequately capture. We specifically compare all models in a panel modeling setting, which provides us with a large sample of 2,086,068 observations. As GB is known to perform exceptionally well with large-scale data, it may be effective in such a panel modeling setting. Therefore, we believe that GB can outperform the simple and widely used HAR-type models. Our original contribution to the literature is threefold: We incorporate EA and offer a detailed analysis; We compare the RV forecasting performance of LightGBM (LGBM) with the simple HAR model; We apply the first and second points using a large dataset of individual stocks in a panel modeling setting, using a slightly different RV measure that incorporates ON information, along with additional feature selection and engineering.

The thesis is organized as follows: Section 2 provides an overview of the related literature and theory. Section 3 elaborates on our hypotheses and methodology. Section 4 outlines our data collection and cleaning procedures, as well as our approach to feature engineering and selection. In Section 5, we present our data analysis. Section 6 describes the models used in our study. Section 7 presents the in-sample results, while Section 8 details the pseudo-OOS results. Finally, Section

9 concludes our thesis.

2 Literature Review and Theory

In this section, we outline a significant portion of the empirical evidence on volatility and provide a brief history and theoretical framework of volatility modeling.

2.1 Empirical Evidence of Volatility

There have been several well-documented empirical observations regarding RV over the years. By RV, we mean a measure of volatility using high-frequency intraday returns as opposed to close-to-close returns or similar measures. Specifically, RV is quantified as the squared sum of all intraday log returns at a chosen sampling frequency. For additional details, please refer to Section (3.2). A comprehensive outline of these findings can be found in the works of Cont (2001) and Masset (2011). These observed statistical properties, known as stylized empirical facts, demonstrate that although asset prices evolve randomly, they consistently share certain statistical characteristics. To summarize, these are: Volatility is non-constant and exhibits long memory; The distribution of volatility is fat-tailed; Shocks are typically followed by aftershocks; Volatility is often negatively correlated with returns; Trading volume shows a correlation with volatility; Volatility measured at different frequencies carries distinct information content.

2.1.1 Long Memory

Despite many simple models assuming constant volatility, the evidence clearly shows that volatility varies over time (Akgiray, 1989; Castanias, 1979; Fama, 1965; Turner & Weigel, 1992). Volatility also exhibits clustering and long memory characteristics. Clustering is a positive correlation between current volatility and its lagged values, a phenomenon also referred to as autocorrelation. Long memory, in essence, indicates that autocorrelation of volatility persists across numerous lags. This particular observation remains remarkably stable across various asset classes and time periods and is considered a typical manifestation of volatility clustering (Teyssière & Kirman, 2007). GARCH models were among the first to capture this phenomenon (Baillie & Bollerslev, 1992; Engle & Rosenberg, 1995). The eventual diminishment of autocorrelation to zero for extended lags provides evidence of mean-reverting volatility (Masset, 2011).

2.1.2 Leptokurtic Distribution

Early evidence of fat tails in financial data can be found in the works of Mandelbrot (1963) and Fama (1965). For a comprehensive overview of research on fat tails in finance, refer to Rachev (2003).

2.1.3 Shocks and Aftershocks

Typically, shocks are followed by aftershocks. H. Liu and Loewenstein (2009) states that the probability of another crash may increase following a crash. Early evidence supporting the clustering of extreme moves is presented in Turner and Weigel (1992).

2.1.4 The "Leverage" Effect

Volatility often exhibits a negative correlation with returns, though the underlying cause of this effect is not fully understood. Two dominant theories have been proposed to explain this phenomenon: the "financial leverage hypothesis" and the "volatility feedback hypothesis". The financial leverage hypothesis posits that when the value of a stock declines, its debt-to-equity ratio rises, which in turn increases the perceived risk of the company and leads to higher return volatility (Black, 1976; Christie, 1982). Conversely, the volatility feedback hypothesis suggests that volatility itself can trigger a risk premium. Accordingly, when stock markets become more volatile, stock prices should decrease to increase expected stock returns (Campbell & Hentschel, 1991; French et al., 1987; Pindyck, 1983; Poterba & Summers, 1984). While the financial leverage hypothesis may explain this phenomenon at a firm level, its applicability falters when considering the phenomenon at an index level. This difficulty in reconciling the theory with empirical observations likely accounts for its limited support (Hibbert et al., 2008). However, empirical evidence appears to favor the financial leverage hypothesis over the volatility feedback hypothesis (Bollerslev, 2006; Masset, 2011). In light of these observations, Hibbert et al. (2008) proposed a new theory based on behavioral arguments. This theory suggests that irrational investors could be the root cause of asymmetry in stock markets.

2.1.5 Volume/Volatility Correlation

It is well established that trading volume exhibits a positive correlation with volatility (Alsubaie & Najand, 2009; Chuang et al., 2009; Karpoff, 1987). Louhichi (2011) found that a significant portion of the volatility persistence can be explained by volume. However, whether volume possesses any predictive power remains uncertain.

2.1.6 Asymmetry in Time Scales

Different frequencies carry different informational content. For instance, low volatility at a longer time horizon is generally followed by low volatility at shorter time horizons. Conversely, high volatility at a long time horizon does not necessarily translate into high volatility at shorter time horizons. This phenomenon, referred to as "asymmetric vertical dependence" (Gençay et al., 2010), is also known as volatility cascade.

2.2 Models and Measures of Volatility

Various models have been proposed in the literature, each with its own benefits and drawbacks. Among model-free measures are historical volatility (HV) and RV, while model-based measures include methods such as GARCH and HAR. Other measures are derived from model-based option prices. In this section, we introduce some of the most significant contributions to the literature.

2.2.1 Historical Volatility

The term "volatility" has been in use for many decades, but its application has grown exponentially with increased computing power and the integration of computers into financial markets. Traditionally, HV was utilized, which is essentially the sample standard deviation as seen in Equation (3). The standard definition of the (unbiased) sample variance is as follows

$$\sigma^2 = \frac{\sum_{t=1}^{T} (r_t - \bar{r})^2}{T-1}, \tag{1}$$

where T represents the number of returns, r_t corresponds to each value of the logarithmic returns, and $\bar{r}$ is the average log return. Log returns are computed as follows

$$r_t = \ln(P_t/P_{t-1}) = \ln(P_t) - \ln(P_{t-1}) = p_t - p_{t-1}. \tag{2}$$

Correspondingly, volatility (the sample standard deviation) is defined as the square root of the variance

$$\sigma = \sqrt{\frac{\sum_{t=1}^{T} (r_t - \bar{r})^2}{T-1}}. \tag{3}$$

Typically, the calculated measure is adjusted to a common period, such as daily, monthly, or yearly (annualized). Assuming independent and identically distributed (iid) returns, variance scales with time while volatility scales with the square root of time. It is a common practice to assume 252 trading days per year. HV is typically calculated using close-to-close (daily) returns before annualizing, to produce a comparable measure across assets and data frequencies, even though the

assumption of iid returns is flawed. While HV is retrospective in nature, the need for more precise and forward-looking estimates of volatility has increased drastically, for reasons explained in the introduction. This necessitates the calculation of implied volatility (IV).

2.2.2 Implied Volatility

Following the work of Black and Scholes (1973), and that of Merton (1973), the well-known Black-Scholes (BS) or Black-Scholes-Merton (BSM) model was proposed. This model is used for pricing European call-and-put options, also known as vanilla options. This model assumes that the asset price follows an exponential/geometric Brownian motion (GBM) described by the following stochastic differential equation (SDE)

$$dS_t = \mu S_t dt + \sigma S_t dW_t. \tag{4}$$

The analytical solution to this SDE, derived using "Itô's formula", is

$$S_t = S_0 \exp\left[\left(\mu - \frac{\sigma^2}{2}\right) t + \sigma W_t\right], \tag{5}$$

which is particularly useful for simulating asset prices. The BS equation

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0, \tag{6}$$

is a parabolic partial differential equation (PDE). Solving this PDE yields the BS formula, an explicit formula for pricing European calls and puts that takes five inputs (six if the asset pays dividends). The value is given by

$$V = f(r, \tau, S, K, \sigma), \tag{7}$$

where $\tau = T - t$ (Hull, 2015). The future RV over the life of the option (σ) is the only parameter that cannot be directly observed in the market. The function f is monotonically increasing in σ, implying that, by the inverse function theorem, there will be at most one value of σ corresponding to a specific value of V. Assuming an inverse function of f, denoted $g = f^{-1}$, we have $\sigma_{\bar{V}} = g(\bar{V}, \cdot)$. Here, $\sigma_{\bar{V}}$ represents the IV derived from the BS formula. While there are numerous other measures of IV, the one obtained from the BS formula is most commonly used. Unless stated otherwise, "IV" refers to BS IV as above. Thus, for options trading in the market, it is possible to infer the IV - a market expectation for the volatility of the underlying over the life of the option. This can be calculated using a root-finding technique like the Newton-Raphson method (see, e.g., Ypma

(1995)) by solving

$$f(\sigma_{\bar{V}}, \cdot) - \bar{V} = 0, \tag{8}$$

(Orlando & Taglialatela, 2017). As such, IV serves as a market-implied (forward-looking) estimate and is widely employed as a measure of expected volatility beyond its application in option pricing. The VIX, a volatility index that reflects the expectation of the stock market regarding volatility in the S&P 500 over the next month (30-day period), is based on S&P 500 index options. Quoted as an annualized standard deviation, it provides an estimate of the implied volatility ("VIX Index", 2023). IV and measures like the VIX are well-recognized for their fairly accurate forecasts of volatility.

2.2.3 Stochastic Volatility

Stochastic volatility (SV) models have been in existence for a considerable amount of time. Some of the earliest papers on SV focused on discrete-time models, were authored by econometricians, and were specifically designed for risk management. Today, SV is almost always associated with continuous-time models, which are more applicable to financial mathematics and options pricing. The key feature of an SV model is that the variance of the process is itself randomly distributed (Shephard, 2005). Some of the more prominent SV models in current use include the Heston model (Heston, 1993), the SABR model (Hagan et al., 2002), and the Hull-White model (Hull & White, 1987). The Heston model is defined as

$$\begin{aligned} dS_t &= \mu S_t dt + \sqrt{\nu_t} S_t dW_t^S \\ d\nu_t &= \kappa(\theta - \nu_t)dt + \xi\sqrt{\nu_t} dW_t^\nu. \end{aligned} \tag{9}$$

This model assumes that the asset follows the stochastic process outlined in the first line, while the instantaneous variance given in the second line follows a Feller square-root process, also known as a CIR process (Cox et al., 1985). This is an extension of the Vasicek model (Vasicek, 1977). The Wiener processes have a correlation coefficient denoted by ρ. The initial variance is represented by ν_0, the long-run variance by θ, the rate at which the instantaneous variance ν_t mean-reverts to θ by κ, and the volatility of volatility (referred to as "vol of vol") by ξ. SV models typically capture features such as volatility clustering and mean reversion.

2.2.4 Local Volatility

Local volatility (LV) models are generalizations of the BS model. Unlike in the BS model, where the volatility is constant, volatility in the LV model is a function

of both time and the asset price (Dupire, 1994),

$$dS_t = r(t)S_t dt + \sigma(S_t, t)S_t dW_t. \tag{10}$$

LV models are useful for options pricing and capturing the entire volatility surface across different maturities and strikes. Derman et al. (1996) states that the BS measure of σ is a sort of global measure of volatility, in contrast to LV, where it depends on a specific state in the model. Subsequent developments include local stochastic volatility (LSV) models (Ren et al., 2007) and path-dependent volatility (PDV) models (Foschi & Pascucci, 2008), which aim to capture the benefits of both LV and SV models. In LSV models, the volatility is modeled as a function of time, asset price, and an additional stochastic process. PDV models, on the other hand, model the volatility based on the asset price path.

2.2.5 ARCH and GARCH

In the early 1980s, the autoregressive conditional heteroscedasticity (ARCH) model was proposed (Engle, 1982). This discrete-time model features time-varying, although deterministic, volatility. ARCH is especially useful when the variance of the error in a time-series model follows an autoregressive (AR) model. AR(p) denotes an AR model of order p (p lags off its own previous values). It is defined as follows

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + u_t, \tag{11}$$

where u_t is white noise with a mean of 0 and constant variance σ^2. ARCH(q) denotes an ARCH model of order q (q lags off its own previous values). It is defined as follows

$$\begin{aligned} y_t &= \beta_0 + \sum_{i=1}^{n} \beta_i x_{i,t} + u_t, \quad u_t \sim \mathcal{N}(0, h_t) \\ h_t &= \alpha_0 + \sum_{i=1}^{q} \alpha_i u_{t-i}^2. \end{aligned} \tag{12}$$

The first equation is the mean equation while the second equation is the variance equation. The mean equation could be specified as an AR, ARMA, or MA model. Here, u_t now follows an AR(q) process with mean 0 and non-constant variance σ_t^2, also called h_t in the literature. ARCH requires $a_i \geq 0 \ \forall i \in 0, \ldots, q$, as the variance has to be non-negative by definition. Engle used ARCH to model the variance of UK inflation. The model significantly improved the forecast compared to a least squares model due to the presence of "ARCH effects" (squared residuals exhibit autocorrelation) in the variance. Hence, ARCH can be effectively used to

forecast volatility, and it quickly gained traction.

Four years later, the generalized ARCH (GARCH) model was introduced (Bollerslev, 1986). It overcomes many of the problems associated with ARCH, such as determining the lag length q (which might be very large) and limiting possible violations of non-negativity constraints. GARCH extends ARCH by allowing dependence on its own previous lags in the variance equation. Thus, GARCH is like an ARMA model for the variance equation. An ARMA model also includes a moving average (MA) part, being lags of the errors. An MA(q) (q lags of its own errors) is defined as follows

$$y_t = \mu + \sum_{i=1}^{q} \theta_i u_{t-i} + u_t, \tag{13}$$

where u_t is white noise with a mean of 0 and constant variance σ^2. ARMA(p, q) denotes an ARMA model of order (p, q) (p lags of its own previous values and q lags of the errors). It is defined as follows

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j u_{t-j} + u_t, \tag{14}$$

where u_t is white noise with a mean of 0 and constant variance σ^2. There are many extensions of ARMA, such as ARIMA and ARFIMA, etc. GARCH(p, q) denotes a GARCH model of order (p, q) (p lags of its own previous values and q lags of the errors). It is defined as follows

$$\begin{aligned} y_t &= \beta_0 + \sum_{i=1}^{n} \beta_i x_{i,t} + u_t, \qquad & u_t &\sim \mathcal{N}(0, \sigma_t^2) \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{q} \alpha_i u_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2. \end{aligned} \tag{15}$$

u_t now follows an ARMA(p, q) process with a mean of 0 and non-constant variance σ_t^2. A GARCH(1,1) can be written as an infinite order ARCH model and is generally sufficient to capture volatility clustering, mean-reversion, and fat tails in the data, although it is symmetric. An example of a GARCH(1,1) model with mean equation AR(1) is an AR(1)-GARCH(1,1) model given by

$$\begin{aligned} y_t &= \mu + \phi y_{t-1} + u_t, \qquad & u_t &\sim \mathcal{N}(0, \sigma_t^2) \\ \sigma_t^2 &= \alpha_0 + \alpha_1 u_{t-1}^2 + \beta_1 \sigma_{t-1}^2. \end{aligned} \tag{16}$$

GARCH is a more parsimonious model than ARCH, thus less risk of overfitting. It is usually estimated by maximum likelihood estimation (MLE). There are many extensions of GARCH, such as GJR-GARCH/TGARCH, AGARCH, and EGARCH (asymmetric and able to capture leverage effects). Alternatively, the errors may be assumed to follow a student-t or skew-t distribution, as opposed to a normal distribution.

2.2.6 SMA and EWMA

Possibly the simplest method for forecasting volatility (other than a naive forecast of the current value or simply predicting the mean) is a rolling or moving average (to be distinguished from moving average models mentioned earlier). Two such models are the simple moving average (SMA) and the exponential moving average (EWMA or EMA), as seen in RiskMetrics (1996). The SMA takes a simple rolling mean (giving equal weights to each observation) over a pre-determined number of periods. SMA for realized volatility (SMA-RV) is defined as follows

$$\sigma_t = \sqrt{\frac{1}{M}\sum_{j=1}^{M} r_{t-j}^2}. \tag{17}$$

With daily log returns, it is common to set $M = 20$ (using the RV over the previous month), and further annualize by multiplying with $\sqrt{252}$. σ_t can be interpreted as an estimate made at $t-1$ of the volatility over the next day, week, or month (a relatively short forecast). It is re-estimated as new observations come in. Each observation has a weight of $1/M$, and for small values of M, the estimate is very sensitive to recent observations. For $M = 1$, the current value is predicted (naive). For $M = T$, the whole sample is used, and the estimate becomes the unconditional volatility.

EWMA gives more weight to newer observations and lesser weight to older ones (using all the observations with exponentially decaying weights, also called smoothing). EWMA for realized volatility (EWMA-RV) is supported by RiskMetrics. It is written recursively as

$$\sigma_t = \sqrt{\lambda\sigma_{t-1}^2 + (1-\lambda)r_{t-1}^2}, \tag{18}$$

assuming an infinite amount of data (which holds approximately for large T). $0 < \lambda < 1$ is the smoothing parameter and is typically set around 0.06. The simple EWMA model does not capture mean-reversion or leverage effects but is generally a fairly accurate forecast for short horizons.

2.2.7 HAR

Subsequently, heterogeneous AR (HAR) models were introduced (Corsi, 2009), specifically termed the HAR model of RV (HAR-RV). This model was designed to additively capture the volatility cascade across various time periods. In simulations, it successfully replicated many empirical properties found in financial data. The parsimonious model is typically used with high-frequency data (HFD). For the context of our paper, we define HFD as data that occurs more frequently than once per trading day. One of its notable features is the capability to capture long memory (persistence) even though the model itself does not possess true long-memory properties. As features, it incorporates RV over different time horizons and is defined as follows

$$RV_{t+1d}^{(d)} = \mu + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + u_{t+1d}. \tag{19}$$

This specifically represents a HAR(3)-RV model due to its three RV terms (hereafter referred to simply as HAR). The HAR model can be easily estimated using OLS and is generally regarded as one of the superior models overall. Several extensions of the HAR model, which capture leverage effects, jumps, etc., have been proposed.

2.2.8 Machine Learning

Recently, more complex machine learning (ML) techniques have gained momentum, although the number of studies remains limited. Varian (2014) discussed the need for ML models to process the ever-increasing big data, which requires enhanced data manipulation tools, such as variable selection due to an increased number of potential predictors, and the capability to handle flexible/complex relationships often accommodated by big data. Some papers have explored regularization and variable selection using Lasso. Audrino and Knaus (2016) challenged the HAR model by testing multiple lags (1-100) and applying a Lasso penalty to identify the most relevant lags through shrinkage. They found that their Lasso-HAR performed equivalently to the HAR model OOS. Caporin and Poli (2017) researched the role of textual data and its impact on volatility forecasting, with a particular focus on news stories (sentiment). Within the penalized regression framework, they found that including such news-related variables improves forecasts. Audrino et al. (2020) conducted a study with a similar focus on Google searches of financial keywords, yielding similar results. Luong and Dokuchaev (2018) utilized the Random Forest (RF) algorithm, documenting improvements with the proposed model on HFD, especially during highly volatile periods.

Mittnik et al. (2015) employed GB (componentwise) for monthly forecasts, including other financial and macroeconomic variables, outperforming commonly used benchmarks including GARCH. They argued that risk drivers affect volatility in a nonlinear fashion and that external economic variables significantly contribute to individual stock volatility. Teller et al. (2022) explored the XGBoost algorithm, demonstrating that it outperforms both HAR and LSTM (Long Short-Term Memory) models for one-step-ahead predictions. For longer horizons, XGBoost with linear base learners outperformed nonlinear specifications, indicating the presence of nonlinearities are more prevalent short-term. Z. Liu (2022) used more than 100 features and built an ensemble model of five ML models (including GB), providing robust and superior OOS R^2 results. Giordani (2021) introduced a new ML model, "boosting of smooth additive regression trees" (SMARTboost), which outperformed XGBoost and OLS across all horizons, particularly at longer horizons as the effective sample size decreases. These forecasts were conducted on global equity indices. Other authors exploring GB include Ding et al. (2022) and Wing-Yi Chio et al. (2022). Several authors have delved into neural networks (NN), a part of deep learning (DL), such as Bucci (2020), Donaldson and Kamstra (1997), Fernandes et al. (2014), Hillebrand and Medeiros (2010) and Rahimikia and Poon (2020). Some of the most recent papers comparing various ML models include Christensen et al. (2021) and Li and Tang (2022).

We wish to highlight the unique aspects of our study: we consider a full-day RV measure that incorporates ON information while aiming to improve forecasts using EA. Furthermore, we have a large panel of stocks that we forecast using both linear HAR models and GB models.

3 Hypotheses and Methodology

In this section, we define our hypotheses and methodology for answering the research question.

3.1 Hypotheses

We investigate the impact of EA and OR within the context of RV forecasting, hypothesizing that the integration of these elements could enhance forecast accuracy. Earnings, known well in advance and typically released in pre-market or after-hours, could augment forecasts, especially in the days surrounding an EA. To more effectively capture this effect, we incorporate the squared ON log return in the RV measure, detailed in Equation (29). We anticipate this new measure will refine the traditional academic RV definition and yield more practically appli-

cable forecasts. Moreover, this measure can function as a comprehensive, full-day measure of RV, enhancing its relevance in options pricing.

To validate our hypotheses, we employ the HAR model and LGBM, applying them to a panel of individual stocks in the S&P 500. We favor GB methods due to their adeptness at capturing potential interaction effects and nonlinearities. We also posit that the integration of large-scale data could significantly improve performance - potentially exponentially so in a GB setting compared to a panel regression setting - given its proficiency with big data. This hypothesis is substantiated by several academic studies, including Bollerslev et al. (2018) and Li and Tang (2022). Consequently, GB might outperform the HAR model.

3.2 Methodology

The first step involves constructing the RV measure for all stocks. We assume that the data-generating process (DGP) is defined by the continuous-time stochastic process as shown in Equation 20, where μ represents the drift term, σ denotes the diffusion, W is a standard BM, and N is the process of the number of jumps with jump size J_i,

$$p_t = p_0 + \int_0^t \mu(s)\mathrm{d}s + \int_0^t \sigma(s)\mathrm{d}W(s) + \sum_{i=1}^{N(t)} J_i. \tag{20}$$

The daily quadratic variation (QV) is given by Equation (21), where the first term is the daily integrated variance (the square root of it is the integrated volatility) and the second term is the jump component (Teller et al., 2022),

$$QV_t^{(d)} = \int_{t-1}^{t} \sigma^2(s)\mathrm{d}s + \sum_{i=N(t-1)+1}^{N(t)} J_i^2. \tag{21}$$

From this, we derive the realized variance estimator as shown in Equation 22, which holds for $m \to \infty$ (Barndorff-Nielsen & Shephard, 2002). Here, t denotes the day defined in one-unit increments while j represents the jth observation of that day, and each day has a frequency of m.

$$\mathrm{RVar}_t^{(d)} = \sum_{j=1}^{m} r_{t,j}^2. \tag{22}$$

Consequently, the jth intraday log return can be represented as follows

$$r_{t,j} = p_{t-1+\frac{j}{m}} - p_{t-1+\frac{j-1}{m}}. \tag{23}$$

We can obtain the RV over different periods such as weekly ($M = 5$), monthly ($M = 20$), etc., by applying the simple moving average (SMA) to each daily RV over the corresponding period, as illustrated in Equation 24, where p denotes the period

$$RV_t^{(p)} = \frac{1}{M}\left(RV_t^{(d)} + RV_{t-1d}^{(d)} + \cdots + RV_{t-(N-1)d}^{(d)}\right). \tag{24}$$

For GB, we utilize a 50/30/20 split for the train, validation, and test samples, whereas, for OLS, an 80/20 division is applied to the train and test samples. This approach ensures that the models are trained on ample data, including the GFC, validated on a substantial sample, and tested on a sizable yet challenging sample that includes the Covid-19 crash. During the training and validation stages, hyperparameter optimization is performed for each model to prevent overfitting and identify the most generalizable model. A limitation, however, is that our training data includes a highly volatile period (GFC), while the validation set does not, which makes forecasting on a test set containing a highly volatile period (Covid-19) more challenging. Nonetheless, we must work with the data available to us, and such information would not be known ex-ante in a real-time forecasting scenario. Initially, we test all models using the same features; that is, when employing the HAR model, we include only the same features in the GB model. Subsequently, both models are augmented with the same additional features. Finally, we retain the benchmark model as is and augment only the GB model for comparison. That is, we will carry out additional feature selection from other sources and feature engineering with the data already collected.

We begin by defining and training all models, exploring Shapley additive explanations (SHAP values) for LGBM on the full sample, whereas for pooled OLS, we present the individual coefficients, t-statistics, and significance levels. SHAP values are considered the best method for measuring feature importance in tree-based boosting models (Lundberg et al., 2018). The hyperparameters we tune using the validation set include the number of leaves (NOL), learning rate (LR), feature fraction (FF), minimum gain to split (MGTS), extra trees (ET), L1 regularization (L1), and L2 regularization (L2). We employ L1 (Mean Absolute Error or MAE) as the loss function to enhance robustness to outliers and mitigate overfitting (compared to, for example, L2 loss or Mean Squared Error (MSE)) and is defined as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|. \tag{25}$$

In fact, we conducted forecasts with L2 loss which yielded poorer results, and Huber loss, which returned similar results to L1 but was more computationally

demanding. These results were consistent for both RMSPE and MAPE.

The last 20% of our data is our "true" (pseudo) OOS test set, containing a "bad" state (Covid-crash). This ensures the models are tested in both a calmer period (pre-Covid and post-Covid) and a turmoil period (intra-Covid). To determine which models perform the best, we compare two measures of model fit, root mean square percentage error (RMSPE)

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}, \tag{26}$$

and mean absolute percentage error (MAPE)

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}. \tag{27}$$

It is important to note that all models are estimated using the logarithm of the RV variables (the convention in the literature), denoted as $\log(RV_t)$. Specifically, we take the logarithm of the three RV variables with horizons $h = 1, 5, 20$.

4 Data Set and Feature Engineering- and -Selection

4.1 Data Collection

The main data set (individual stock prices) was obtained from FirstRate Data and is known as the S&P 500 Historical Intraday Prices Bundle (FirstRate Data, 2023). It includes open, high, low, close, and volume (OHLCV) data, with volume measured in terms of the number of shares traded, spanning from January 3, 2005, to October 21, 2022. The data is available in 1-minute, 5-minute, 15-minute, 30-minute, and 1-hour intervals. Our data set only includes bars when there is trading volume, so any gaps in the bars can be attributed to periods with zero volumes traded. All data has been adjusted for dividends and splits, although complete information on dividends (date and dividend amount) and splits (date and split ratio) is included. For future reference, we have defined the time periods as illustrated in Table (1).

Time-period definition	Acronym	Start time - End time	Hours
Pre-market	PM	04:00-08:00	4.0
Normal pre-market	NPM	08:00-09:30	1.5
Regular hours	RH	09:30-16:00	6.5
After-hours	AH	16:00-20:00	4.0
Full trading day	FTD	04:00-20:00	16.0
No trading	NT	20:00-04:00	8.0
Overnight	ON	16:00-09:30	17.5

Table 1: Our definition of the different time periods.

Furthermore, we define PM and AH as extended hours, which are also included in the dataset. PM and NPM can alternatively be referred to as PM alone (04:00-09:30). The OR is typically defined as the close-to-open return, while the RH return corresponds to the open-to-close return (and the daily return corresponds to the close-to-close return). The data set encompasses the component stocks of the S&P 500 as of September 21, 2022, as well as those previously included in the S&P 500 that were in existence during the sample time frame of the data.

We acquired daily OHLC values for the VIX index, VVIX index, MOVE index, S&P 500 index, the US Dollar index, and Crude Oil for the same period from Yahoo Finance (2023). Much like the VIX index, the MOVE index represents a measure of volatility in the US bond market, specifically for Treasuries. The VVIX index tracks the volatility of the VIX index itself, effectively measuring the vol of vol on the S&P 500. The US Dollar index (DXY) gauges the value of the US Dollar relative to a select group of foreign currencies, which include (in descending order of weight) the EUR, JPY, GBP, CAD, SEK, and CHF. Additionally, we obtained the earnings dates for all US stocks filed with the SEC dating back to 1994 from EarningsDates (2023).

4.2 Data Cleaning

We were initially provided with a total of 668 unique stocks. To ensure robustness, we excluded any stocks that had fewer than 3750 days of intraday HFD, which is equivalent to approximately 15 years of data, assuming 250 trading days per year. This left us with a selection of 499 stocks. Further, when we limited our data to only those stocks for which we have earnings announcement (EA) data, our sample size was reduced to 478 stocks.

In the absence of microstructure noise, it would be optimal to sample as fre-

quently as possible. However, in the presence of microstructure noise, increasing the frequency leads to a decrease in bias but an increase in variance. Consequently, it becomes necessary to optimize the bias-variance trade-off (Aït-Sahalia et al., 2005; Andersen et al., 2011). Several studies suggest that using a subsampling and averaging procedure is optimal. Yet, our approach deviated from this recommendation; we utilized the 5-minute data acquired to build all features. This method aligns with most of the relevant literature (Andersen et al., 2011; Ghysels & Sinko, 2011; L. Y. Liu et al., 2015). Shorter data sampling frequencies, such as 1-minute, may be too noisy due to microstructure noise, while longer data sampling frequencies, like 1 hour, may be too imprecise for optimal forecasting performance. During RH (from 09:30 to 16:00), there will be $12 \times 6.5 = 78$ 5-minute observations used to construct the daily RV. The daily RH RV can be computed by squaring Equation (22)

$$RV_t^{(RH,d)} = \sqrt{\sum_{j=1}^{m=78} r_{t,j}^2}. \tag{28}$$

Our methodology diverges from the traditional approach. The academic definition of $RV_t^{(d)}$ ($RV_t^{(RH,d)}$) focuses exclusively on the volatility that occurs during RH, thereby neglecting the full-day volatility. By slightly modifying this measure to include ON fluctuations, we can construct a more comprehensive measure of daily RV. This new measure is defined as follows

$$RV_t^{(RHON,d)} = \sqrt{\sum_{j=1}^{m=78} r_{t,j}^2 + on_t^2}. \tag{29}$$

In this context, $on_t^2 = \log\left(1 + (P_t^{(o)}/P_{t-1}^{(c)} - 1)\right)^2$ represents the square of the ON log return (specifically, the square of the close-to-open log return), where $P_t^{(o)}$ is the opening price on day t and $P_{t-1}^{(c)}$ is the closing price on day $t-1$. We introduce a new term, RHON (RH plus ON), to denote this modified period utilized to build the RV measure. This measure has been explored by authors such as Bollerslev et al. (2018) and Hansen and Lunde (2005). Alternatively, one could consider the FTD dynamics, from PM, through RH, to AH (i.e., 04:00-20:00), an approach which, to our knowledge, has not yet been researched. This does omit potential trading in dark pools outside these hours, but such trading is likely minimal and often executed for non-informational reasons (noise). For FTD, there would theoretically be $12 \times 16 = 192$ such observations. However, due to limited trading activity in the extended hours for many stocks, the actual number of 5-minute observations utilized to build the daily RV measure would likely be fewer, but still

higher than those for RH and RHON. We define this new measure as follows

$$RV_t^{(FTD,d)} = \sqrt{\sum_{j=1}^{m=192} r_{t,j}^2}. \tag{30}$$

For all measures, we calculate the weekly (w) and monthly (m) RV by following the usual SMA approach as in Equation (24).

4.3 Feature Engineering

We began by constructing all the volatility measures as defined above (RH, RHON, and FTD) for $h = 1, 5, 20$, resulting in a total of nine distinct variables. Additionally, we computed the same types of features using the absolute value in lieu of squared log returns, as described in Forsberg and Ghysels (2007). According to several statistical tests, they found that absolute returns correlate more strongly with volatility than returns do, a trend observed across various assets and time periods. The primary advantage of absolute returns, as they argue, is their resistance to outliers. Instead of using the term "RV" for these measures, we denote them as "RAV" (realized absolute volatility). Subsequently, we turned our attention to the NPM volatility and calculated both RV and RAV based on data from NPM (08:00-09:30) for $h = 1, 5, 20$. Empirical studies suggest that PM volatility possesses predictive power for next-day volatility (C.-H. Chen et al., 2012; Zhu et al., 2017). As we are employing the L-HAR model as our benchmark, we computed truncated log returns as follows

$$r_t^{(p)-} = \min\left(0, r_t^{(p)}\right), \tag{31}$$

where $r_t^{(p)}$ is the mean log return over the period, calculated as $r_t^{(p)} = \frac{1}{h}\sum_{i=1}^{h} r_{t-i+1}$, for $h = 1, 5, 20$. The L-HAR model is further described in Section (6.1).

We calculate 1-day ON returns (non-log) as an additional feature, using close-to-open prices. Moreover, we truncate these returns using both the min and the max operator, to obtain three distinct ON return features (the "symmetric"/regular return, the "negative part", and the "positive part"). Similarly, during RH, we incorporate a feature containing the RH symmetric return and another for the RH positive part of the return.

Following the methodology outlined in Mei et al. (2017) we construct measures

for realized kurtosis (RK)

$$RK_t^{(d)} = \frac{m \sum_{j=1}^{m} r_{t,j}^4}{(RV_t^2)^2}, \tag{32}$$

and realized skewness (RS)

$$RS_t^{(d)} = \frac{\sqrt{m} \sum_{j=1}^{m} r_{t,j}^3}{(RV_t^2)^{3/2}}. \tag{33}$$

The authors ascertain that both RK and RS significantly and negatively impact future volatility. Neither is beneficial for short-term forecasting; however, both enhance forecasting at mid- and long-term horizons, with RS proving more useful. Consequently, we have incorporated features only for $h = 5$ and $h = 20$, employing the SMA approach as previously described.

Furthermore, we attempt to capture the changes in RV. To accomplish this, we calculate the realized quarticity (RQ) as

$$RQ_t^{(d)} = \frac{m}{3} \sum_{j=1}^{m} r_{t,j}^4, \tag{34}$$

which serves as a measure of 1-day vol of vol (Corsi et al., 2008). For $h = 5$ and $h = 20$, we adopt a different approach, calculating the rolling h-day standard deviation of the RV. Consequently, we obtain three vol of vol measures, one for each horizon. To gauge the daily (Dollar) volume, we construct a feature by multiplying the daily volume during RH with the volume weighted average price (VWAP) during the same period.

Additionally, we construct three features to capture the short-term mean, medium-term mean, and long-term mean, utilizing an exponentially weighted moving average (EWMA) of the 1-day RV during RH. We apply different decay rates for each feature, corresponding to an effective sample size (center-of-mass or CoM) of 50, 250, and 1,250 trading days, respectively. The computation is done as follows

$$\operatorname{Exp} RV_t^{\operatorname{CoM}(\lambda)} = \sum_{i=1}^{h} \frac{e^{-i\lambda}}{e^{-\lambda} + e^{-2\lambda} + \ldots + e^{-h\lambda}} RV_{t+1-i}, \tag{35}$$

where $\operatorname{CoM}(\lambda) = e^{-\lambda} / \left(1 - e^{-\lambda}\right)$ and λ is the decay rate, defined as $\lambda = \log(1 + 1/\operatorname{CoM})$. For an extensive explanation and application, see Bollerslev et al. (2018).

Lastly, given our focus on EA and its impact on RV in a separate discussion of this thesis, it was necessary to introduce some indicator variables (dummies) to capture this effect. To achieve this, we incorporated a dummy variable for the day of EA, one and two days prior to EA, as well as one and two days following EA. As $h = 5$ and $h = 20$ will not significantly utilize any of these dummy variables, we created additional features to allow the model to anticipate, 5 and 20 days prior to EA, that an EA will fall within the prediction horizon.

4.4 Feature Selection

As previously mentioned regarding data collection, we obtained supplementary data from Yahoo Finance. With the VIX and VVIX, we aim to gauge the volatility of the S&P 500 and the volatility of the VIX itself, respectively. These can be viewed as "global" risk measures. According to the "Risk Everywhere" paper (Bollerslev et al., 2018), the authors demonstrate that such global risk factors encompass information not entirely captured by the asset-specific features. Through the MOVE index, we intend to represent the volatility within the crucial US Treasury market. We incorporate the daily closing values of VIX, VVIX, and MOVE as features, as well as the 1-day, 5-day, and 20-day percentage changes in these indices. Reinforcing the Risk Everywhere argument, we calculate the 1-day, 5-day, and 20-day volatility of the Crude Oil price, the Dollar index (DXY), and the S&P 500 index utilizing daily data. To our knowledge, no authors have included MOVE, VVIX, or DXY in RV forecasting, although the authors of Risk Everywhere included a feature similar to MOVE (US 10-year) and DXY (USD/EUR). With daily data, calculating intraday RV is not feasible, necessitating a different approach. Range-based estimators are common, relying on measures such as the high and low prices of a given day (Alizadeh et al., 2002). We compute this as follows

$$RV_t^{LDR} = \log(\text{High}_t) - \log(\text{Low}_t), \tag{36}$$

In this equation, the log difference between the high and low price of the day (log daily range or LDR) serves as our measure of volatility.

Finally, we include the closing level of the S&P 500 index, the 1-day return of the index, and the drawdown of the index as additional features. The drawdown of the index was calculated using the following equation

$$DD_t = \min\left(0, \frac{V_t - \max\left(V_{1:t}\right)}{\max\left(V_{1:t}\right)}\right), \tag{37}$$

where V denotes the index level.

Our finalized panel data set comprises 2,086,068 daily observations (rows) spanning 76 features (columns), amounting to a total of 158,541,168 data points. The original dataset was substantially larger - roughly 78 times the current row count, or 162,713,304 rows, totaling 12,691,637,712 data points - as we transformed 5-minute data into daily measures. To manage each stock in the panel, we devised two features: "idcat" (categorical ID using the ticker) and "idval" (numeric ID). Depending on the model employed, one of these was utilized in the panel for forecasting.

5 Data Analysis

5.1 Volatility Measure

Figure (1) illustrates how our proposed new measure of volatility $RV_t^{(RHON,d)}$ behaves differently compared to the traditional academic definition $RV_t^{(RH,d)}$, and the FTD definition $RV_t^{(FTD,d)}$. A larger portion of the distribution is situated at higher values of 1-day RV as compared to RH, though it remains lower than FTD. This is not surprising, as we incorporate an additional squared log return into the traditional measure, which, by definition, will exceed RH (unless the ON log return is 0). We posit that this measure is more capable of capturing daily actual uncertainty, as standard RV measures tend to underestimate volatility by excluding close-to-open information. Furthermore, we hypothesize that a model employing the new measure will yield greater benefits from the inclusion of EA dummy variables, as earnings are typically released in PM or AH. The RHON measure, in this case, can effectively capture such ON return volatility.

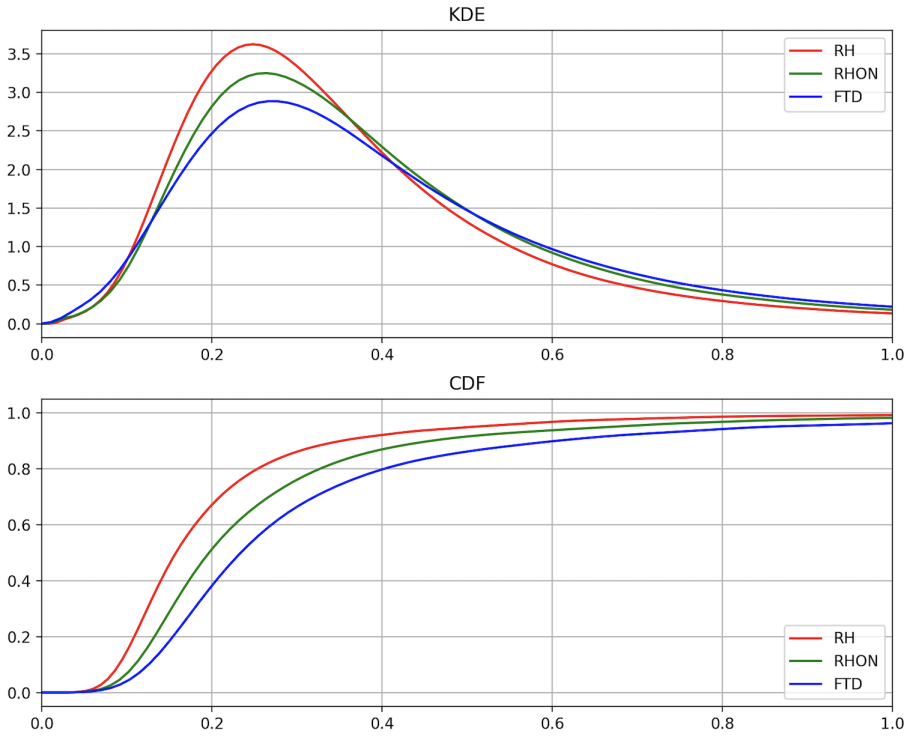

Figure 1: Kernel Distribution and Cumulative Density Function of 1-day RV for three measures of volatility. x-axis truncated at 100% RV. Average values of the S&P 500 stocks.

Although we do not undertake a detailed examination of the FTD measure in this study, we perceive the potential for future analysis. One primary drawback is that the measure combines two disparate processes: RH trading and extended-hour trading, which could potentially be quite different. Regardless of this theoretical shortcoming, empirical evidence may determine its potential for RV forecasting, particularly in attempting to capture the full daily distribution of RV.

5.2 Earnings Announcements

As shown in Figures (2) and (3), RV around EA experiences a systematic increase a few days prior to the announcement, peaks on the day of the announcement, and gradually decreases afterward. The rate of decrease post-announcement is slower than the increase preceding it, remaining elevated for over 20 days post-announcement. All volatility measures show a notable increase in mean (median) volatility of approximately 13-16pp (6-9pp), or about 40-50% (20-40%).

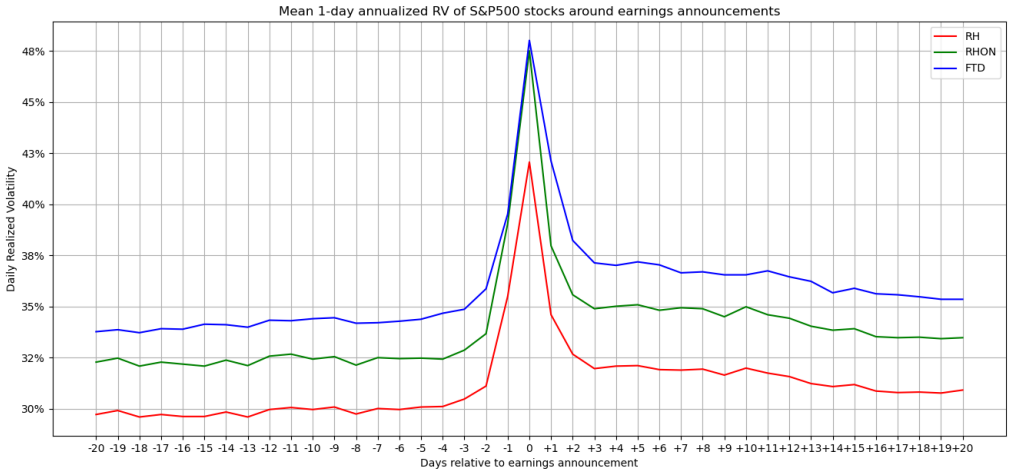

Figure 2: EA and its impact on mean 1-day annualized RV of the S&P 500 stocks. Illustrated for three different measures of volatility, from 20 days prior to the EA until 20 days after the EA.

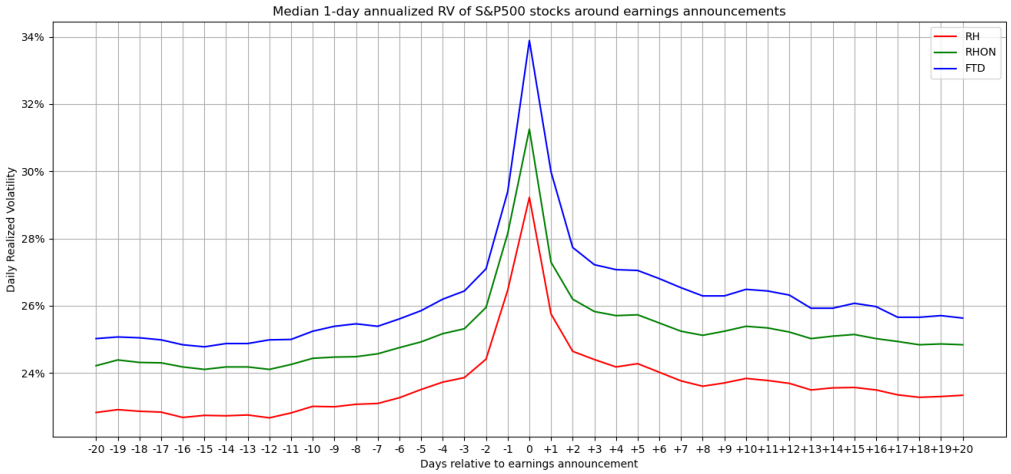

Figure 3: EA and its impact on median 1-day annualized RV of the S&P 500 stocks. Illustrated for three different measures of volatility, from 20 days prior to the EA until 20 days after the EA.

We found it intriguing that the RH measure displayed substantial movement, as illustrated in Figures (2) and (3). While some researchers forecasting RV omit the first and last minutes of each trading day (refer to Stoll and Whaley (1990)), we opted not to adopt this methodology. This led us to question if this could be attributed to highly volatile trading during the first and last minutes of trading days surrounding EA. To investigate this, we formulated another RH measure, "RHcut", which excludes the first and last 10 minutes of trading. Intriguingly, the RHcut measure still exhibits similar levels of movement. As a result, we infer that this could be a consequence of a shift in prices that prompts a volatility surge

from close to open (ON), with high volatility persisting throughout the trading day following the release. Refer to Figure (4).

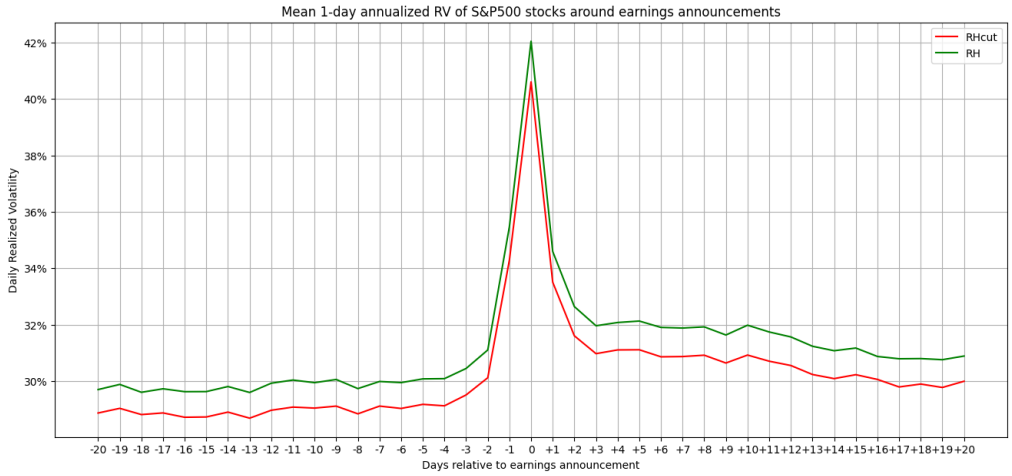

Figure 4: EA and its impact on mean 1-day annualized RV of the S&P 500 stocks. Illustrated for two different measures of RV during RH ("RHcut" excluding the first and last 10 minutes of trading and RH), from 20 days prior to the EA until 20 days after the EA.

Observing Figures (5) and (6), a similar pattern is exhibited for both high-volatility (HV) and low-volatility (LV) stocks. Nevertheless, HV stocks appear to demonstrate a more significant percentage increase on EA days compared to LV stocks, suggesting the volatility of RV is substantially higher for HV stocks. Furthermore, LV stocks commence their rise earlier and maintain heightened levels for a longer duration. Our conjecture is that HV stocks often share some of the following characteristics: naturally uncertain, complex to analyze, and lower liquidity. This could lead to more pronounced upside or downside surprises on EA days, and higher trading volume due to increased buying or selling pressure. Interestingly, the mean 1-day annualized RV appears greater for the RHON measure than the FTD measure for HV stocks. We posit that this could be attributed to frequent earnings surprises in HV stocks and the construction of the RHON measure, which includes a 17.5-hour ON return window, in contrast to multiple smaller return windows in the FTD measure.

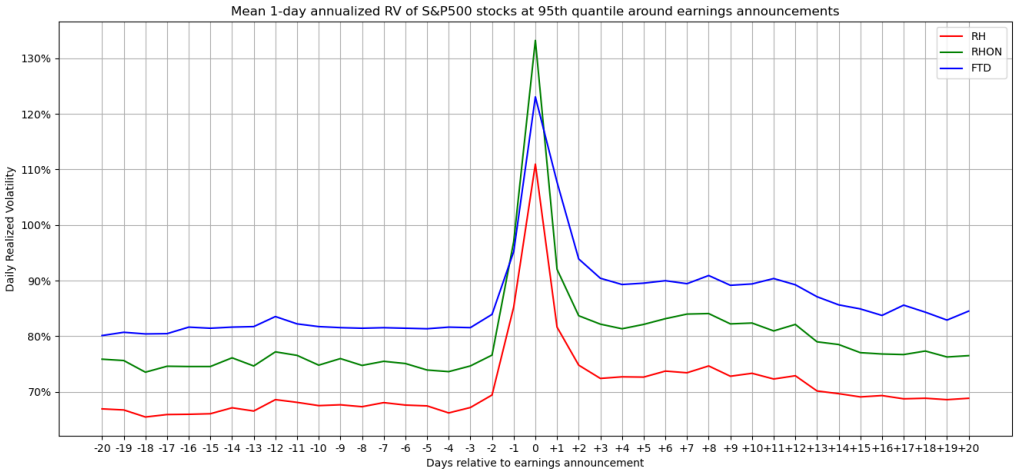

Figure 5: EA and its impact on 95th quantile 1-day annualized RV of the S&P 500 stocks. Illustrated for three different measures of volatility, from 20 days prior to the EA until 20 days after the EA.

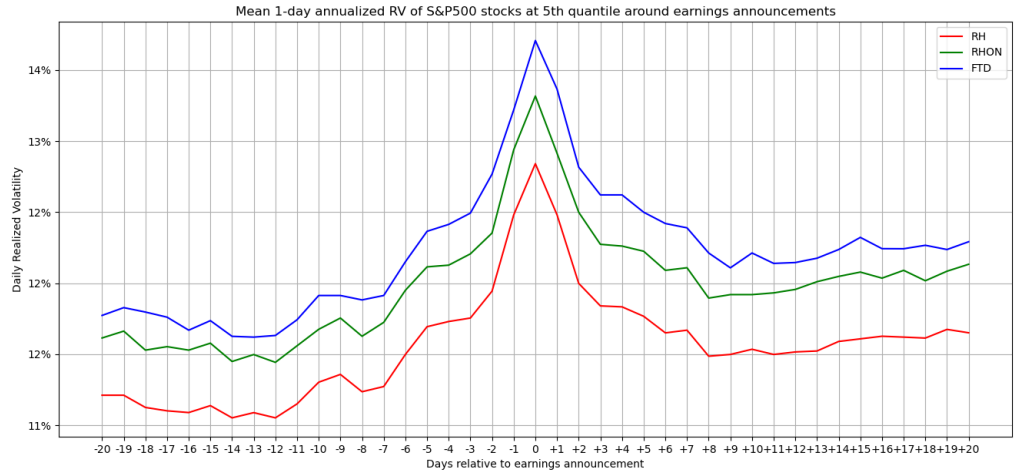

Figure 6: EA and its impact on 5th quantile 1-day annualized RV of the S&P 500 stocks. Illustrated for three different measures of volatility, from 20 days prior to the EA until 20 days after the EA.

One might hypothesize that the mean or median 1-day RV could be higher on days of EA, yet, due to certain factors, it might be lower than the mean or median over longer time horizons (that is, exhibiting low volatility in the days preceding the EA). However, upon examination of the mean and median RV as displayed in Figure (7), it becomes apparent that volatility is elevated on EA days across all time horizons.

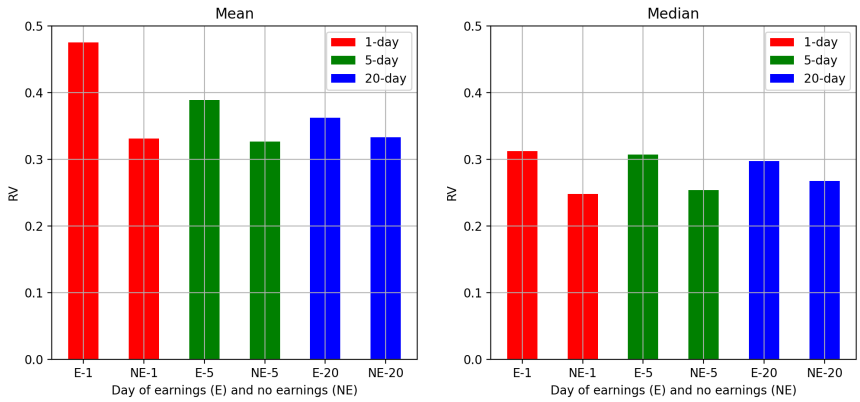

Figure 7: Mean and median RV of the S&P 500 stocks. Illustrated at the different time horizons using our proposed measure of volatility, RHON.

Interestingly, Figures (8) and (9) demonstrate significant disparities among individual stocks. Our interpretation is that some stocks benefit more when controlling for EA, especially at the shorter horizons. Yet, even at the 20-day horizon, the mean value for most stocks appears to be influenced. Our analysis reveals that the mean (median) RV in RHON for days with EA is higher for 83% (85%) of stocks compared to non-EA days at the 1-day horizon, 85% (86%) at the 5-day horizon, and 91% (92%) at the 20-day horizon.

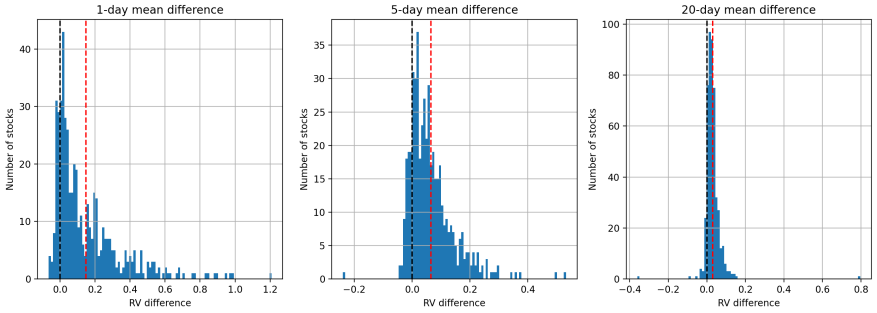

Figure 8: Differences in the mean RV on days of EA versus non-EA days among the S&P 500 stocks are illustrated at the different time horizons utilizing our proposed measure of volatility, RHON. The red dotted line represents the mean of these differences, while the black dotted line signifies zero mean difference, suggesting that volatility remains unaffected by EA.

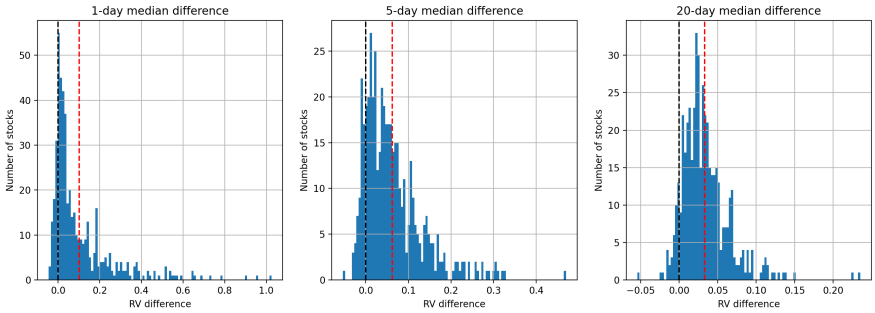

Figure 9: Differences in the median RV on days of EA versus non-EA days among the S&P 500 stocks are illustrated at the different time horizons utilizing our proposed measure of volatility, RHON. The red dotted line represents the mean of these differences, while the black dotted line signifies zero mean difference, suggesting that volatility remains unaffected by EA.

Table (2), (3), and (4) detail regression statistics on EA at the 1-day, 5-day, and 20-day horizons, respectively. As anticipated, the variables "earn" (day of EA), "1bef" and "2bef" (one and two days before EA) all display positive coefficients and are statistically significant. The variables "1aft" and "2aft" (one and two days after EA) have negative coefficients and are also significant. This aligns with the previous illustrations of the impact of EA on RV. Unsurprisingly, "earn" has the highest t-statistic and the highest coefficient. The coefficient of 0.1996 implies that, on days of EA, RV increases by approximately 20% compared to non-EA days, when holding all other variables constant. At the 5- and 20-day horizon, we still see positive coefficients that are highly significant, suggesting that including these variables is likely to adjust the RV prediction upwards.

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
earn	0.1996	15.5541	0.0000	0.0128	0.1744	0.2247
1bef	0.0627	8.1040	0.0000	0.0077	0.0475	0.0778
2bef	0.0212	3.0997	0.0019	0.0068	0.0078	0.0346
1aft	-0.0169	-2.1398	0.0324	0.0079	-0.0323	-0.0014
2aft	-0.0882	-10.8117	0.0000	0.0082	-0.1042	-0.0722

Table 2: Pooled OLS regression results on EA features at the 1-day horizon. Using the full sample and our RHON measure with Driscoll-Kraay (HAC or Newey-West type correction for panel data) covariance estimator. Shown for 95% confidence interval. Estimated using the EL-HAR model described in Equation (39).

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
earn5	0.1324	12.2841	0.0000	0.0108	0.1113	0.1536

Table 3: Pooled OLS regression results on EA features at the 5-day horizon. Using the full sample and our RHON measure with Driscoll-Kraay (HAC or Newey-West type correction for panel data) covariance estimator. Shown for 95% confidence interval. Estimated using the EL-HAR model described in Equation (39).

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
earn20	0.0935	8.8014	0.0000	0.0106	0.0727	0.1143

Table 4: Pooled OLS regression results on EA features at the 20-day horizon. Using the full sample and our RHON measure with Driscoll-Kraay (HAC or Newey-West type correction for panel data) covariance estimator. Shown for 95% confidence interval. Estimated using the EL-HAR model described in Equation (39).

6 Models

6.1 OLS

The traditional OLS-based HAR model by Corsi, which is both simple and accurate, effectively captures the long memory in volatility data - refer to Equation (19). The extension that includes leverage effects by Corsi and Renò (2012) (known as the leverage HAR or L-HAR model), is widely acknowledged as one of the best-performing models. We use it as our benchmark model, which is defined as follows

$$RV_{t+h}^{(\text{L-HAR},h)} = \mu + \sum_{h=1,5,20} \beta^{(h)} RV_t^{(h)} + \sum_{h=1,5,20} \gamma^{(h)} r_t^{(h)-} + u_{t+h}. \tag{38}$$

Next, we augment the L-HAR model with EA to create the "EL-HAR" model

$$RV_{t+h}^{(\text{EL-HAR},h)} = RV_{t+h}^{(\text{L-HAR},h)} + \psi^{(h)} EA_t^{(h)}. \tag{39}$$

For $h = 1$ we include a total of five dummies: one for the day of the EA, one for each of the two preceding days to the EA, and one for each of the two following days after the EA. However, for $h = 5$ ($h = 20$) we only include one dummy each, containing ones for the four (19) days leading up to the EA as well as the day of EA. Additionally, we refer to the model that follows the structure of the EL-HAR model but incorporates RAV features instead of the typical RV features as "EL-HAR-RAV".

We further extend the EL-HAR model to include the short, medium, and long-term exponential mean RV, which we term "EL-HAR-ExpSML" (ExpSML stands for exponential short, medium, and long-term mean). The model is defined as follows

$$RV_{t+h}^{(\text{EL-HAR-ExpSML},h)} = RV_{t+h}^{(\text{EL-HAR},h)} + \sum_{h=50,250,1250} \lambda^{(h)} \text{ExpMean}_t^{(h)}. \tag{40}$$

Next, we experiment with the EL-HAR-ExpSML model extended with closing values and 5-day percentage changes in VIX, VVIX, and MOVE. This extended model is denoted as "EL-HAR-ExpSML-RE" (RE stands for risk everywhere). It is defined as follows

$$\begin{aligned} RV_{t+h}^{(\text{EL-HAR-ExpSML-RE},h)} &= RV_{t+h}^{(\text{EL-HAR-ExpSML},h)} + \beta^{(1)}\text{VIX}_t \\ &+ \beta^{(2)}\text{VVIX}_t + \beta^{(3)}\text{MOVE}_t \\ &+ \beta^{(4)}\text{VIX}_t^{\text{5d chg.}} + \beta^{(5)}\text{VVIX}_t^{\text{5d chg.}} + \beta^{(6)}\text{MOVE}_t^{\text{5d chg.}}. \end{aligned} \tag{41}$$

Finally, "Model X" is used to refer to the model that incorporates all of the features discussed in Section (4).

6.2 Gradient Boosting

6.2.1 Boosting and Gradient Boosting in General

Two methods for aggregating predictions from different models are bagging (bootstrap aggregating) (Breiman, 1996) and boosting (Freund & Schapire, 1997). While boosting combines predictions from multiple models, bagging combines predictions from several instances of the same model. Bagging decreases variation and aids in avoiding overfitting by averaging multiple predictions. It does not, however, resolve issues related to bias. In contrast, boosting prioritizes the models that perform better on the training data in order to reduce bias by iteratively modifying the weights (weighted average) of several models.

The bias-variance trade-off can be decomposed as follows (Hastie et al., 2009). One can estimate $\hat{f}(X)$ of $f(X)$, with Y as the target variable and X as the features. The squared prediction error can be expressed as

$$\begin{aligned} \text{Error}(x) &= E\left[(Y - \hat{f}(x))^2\right] \\ &= (E[\hat{f}(x)] - f(x))^2 + E\left[(\hat{f}(x) - E[\hat{f}(x)])^2\right] + \sigma_e^2 \\ &= \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}. \end{aligned} \tag{42}$$

This expression assumes there is a relationship $Y = f(X) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$. The model cannot reduce the irreducible error. Ideally, given the true model and infinite data, the bias and variance could be reduced to zero. However, in practice, the focus lies on optimizing this bias-variance trade-off.

The foundational principle behind GB is to repeatedly train a series of models, with each one aiming to address the shortcomings of the previous (Friedman, 2001). At each iteration, the gradient (i.e., the slope of the error function) of the loss function $L(y, f(x))$ with respect to the previous predictions of the model is calculated. The new model is then fitted with this gradient and added to the ensemble of models. Gradient-boosted decision trees (GBDT) provide a practical approach that can be employed for both classification (GBCT) and regression (GBRT) problems. In this thesis, we will be using GBRT. Some of the loss functions for regression include squared error (L2), absolute error (L1), and Huber. Both L1 and Huber are more robust to outliers.

GB proves especially beneficial when dealing with high-dimensional data, nonlinearities, and interaction effects. The algorithm operates by iteratively adding weak learners to the model. Each learner is trained to anticipate the residual errors of the preceding model. This process gradually reduces the errors of the model over multiple rounds, thereby enhancing its overall performance.

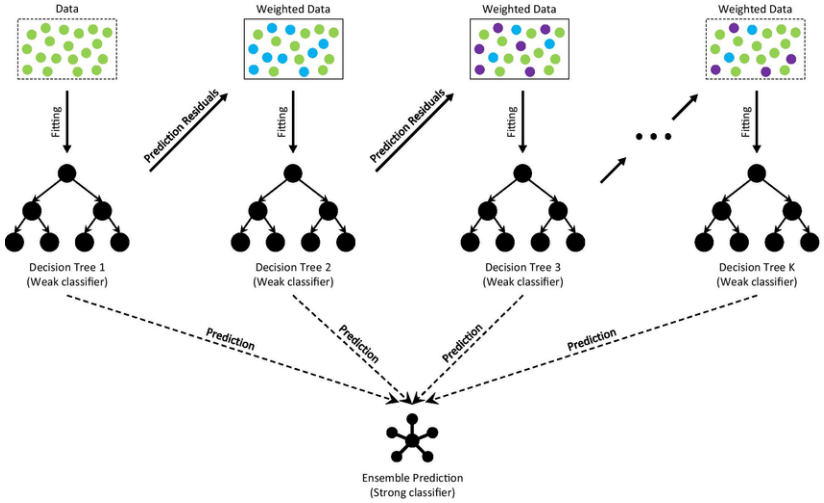

Figure 10: GB architecture (Deng et al., 2021).

One of the primary benefits of GB is its adaptability. GB does not perform feature selection directly, but less relevant features tend to be naturally de-emphasized over iterations in the weak learners. The frequency with which a feature is used in the ensemble and the subsequent reduction in loss typically indicates its relevance. Given the complexity of the model and the nonlinear relations between the

features and the target variable, GB might be more challenging to comprehend than HAR models. However, techniques such as feature importance analysis and partial dependence plots can help in understanding and interpreting the behavior of the model (Greenwell et al., 2018). GB can approximate nonlinear relationships between features and the target variable, as any arbitrary function can be used as a model in the ensemble. Hence, GB can handle complex data comprising nonlinear patterns. GB can also model feature interactions, where the impact of one feature on the target variable depends on the value of another feature. This is accomplished by integrating feature-feature interactions in the weak learner models. GB is capable of processing inputs with both continuous and discrete variables. Weak learner models can use regression trees to divide the feature space for continuous variables, while decision trees or other techniques can be employed for splits depending on the categories of discrete variables. Nevertheless, there are some potential drawbacks worth noting. The algorithm can overfit the training data due to several reasons such as using too many iterations (number of trees), not employing a validation set, a complex model (high number of features), high tree depth, or if the learning rate is set too high. Also, it can be time-consuming and computationally expensive, particularly for large datasets. For additional information, see Bentéjac et al. (2021) and Hastie et al. (2009).

6.2.2 Light Gradient Boosting Machine (LightGBM)

LGBM is a GB algorithm developed by Microsoft in 2016 (Ke et al., 2017). It shares many benefits with Xtreme Gradient Boosting (XGBoost), which was developed in 2014 (T. Chen & Guestrin, 2016), but it offers improved speed with little to no compromise on accuracy (Bentéjac et al., 2021). LGBM employs a leaf-wise (or best-first) approach to growing its trees. See Figure (11).

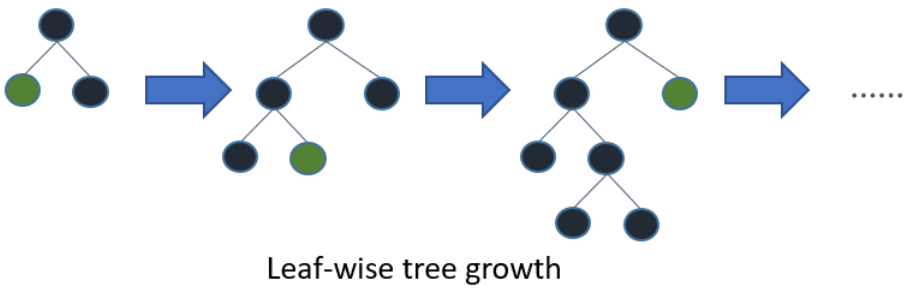

Figure 11: LGBM tree growth representation (LightGBM, 2023a).

In contrast, XGBoost uses a level-wise (or depth-wise) approach for tree growth, as illustrated in Figure (12).

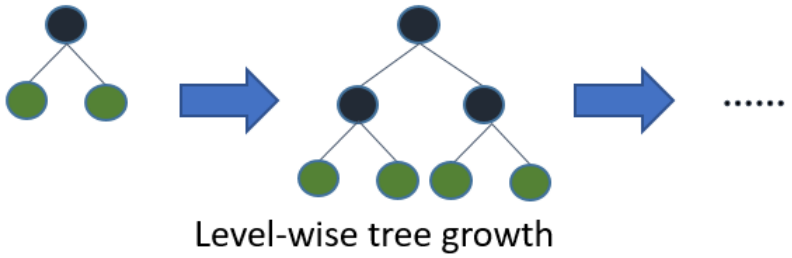

Figure 12: XGBoost tree growth representation (LightGBM, 2023b).

7 In-Sample Results

This section presents the feature importance and Shapley Additive Explanations (SHAP values) on the full sample for LGBM, along with the individual coefficients, t-stats, and significance levels for pooled OLS. SHAP values are widely acknowledged as the optimal method for measuring feature importance in tree-based boosting models (Lundberg et al., 2018). In the following SHAP plots, features are ordered according to their importance, from highest (top) to lowest (bottom). Due to space constraints, only the top 20 features are presented. This section presents only the values for $h = 20$, whereas the values for $h = 1$ and $h = 5$ are included in the appendix. All comments in this section are based on the SHAP values. Pooled OLS results are ranked from lowest to highest p-value. Diligent readers may interpret the pooled OLS results independently.

The feature ranking of the L-HAR model in Figure (13) appears realistic, with the RV lags ranking as the top features. As anticipated, higher values of the RV lags increase the prediction, while lower values decrease it. In line with empirical evidence, lower values of leverage features increase the prediction. Higher values of leverage features seem to slightly reduce the forecasted RV. Upon introducing EA as seen in Figure (14), the impact of leverage features on the model output diminishes, and the importance of the EA feature surpasses them. The EL-HAR-RAV model, as illustrated in Figure (15), displays results very similar to those of the EL-HAR, except that the RAV feature at $h = 5$ now possesses the greatest feature importance compared to $h = 20$ for the RV feature in the EL-HAR model. Notable differences are observed in the EL-HAR-ExpSML model upon adding expanding means, see Figure (16). The short-term expanding mean (50 days) emerges as the most important feature. The medium-term expanding mean also holds substantial importance, ranking above the EA feature. In contrast, the long-term expanding mean appears to hold lesser importance. In the EL-HAR-ExpSML-RE model, there are no significant differences, although some of

the "risk everywhere" features climb above the leverage features, among others. See Figure (17) for an illustration. Model X reveals an interesting observation, as seen in Figure (18). Notably, the EA feature now ranks as the second most important feature. Aside from this, the feature ranking remains fairly consistent, with no major surprises.

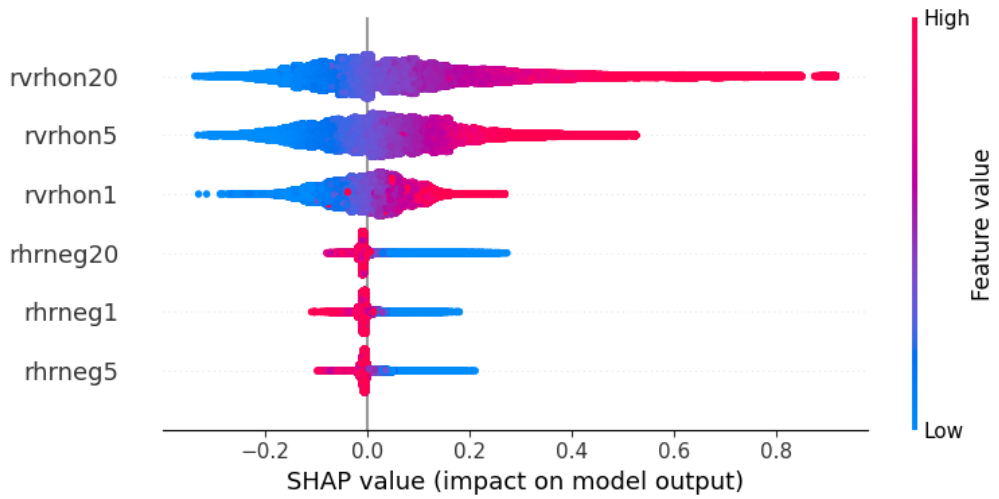

Figure 13: L-HAR: LGBM SHAP values at $h = 20$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
rvrhon1	0.1587	22.4084	0.0000	0.0071	0.1448	0.1726
rvrhon5	0.1640	9.2171	0.0000	0.0178	0.1291	0.1989
rvrhon20	0.4373	21.5808	0.0000	0.0203	0.3976	0.4771
rhrneg1	-1.2421	-11.0078	0.0000	0.1128	-1.4632	-1.0209
rhrneg20	-7.5614	-6.3637	0.0000	1.1882	-9.8903	-5.2326
rhrneg5	-3.4150	-6.2883	0.0000	0.5431	-4.4794	-2.3506

Table 5: L-HAR: Pooled OLS results at $h = 20$

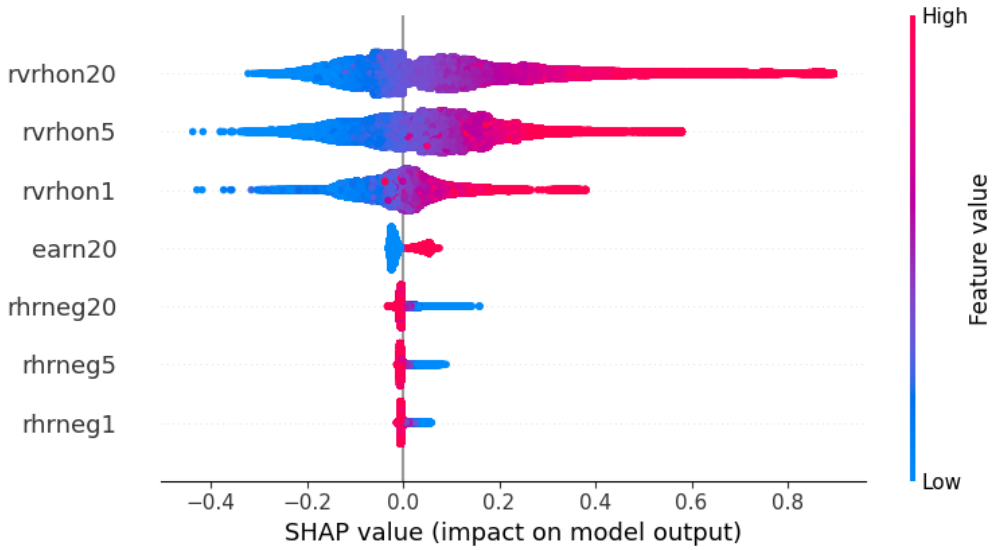

Figure 14: EL-HAR: LGBM SHAP values at $h = 20$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
rvrhon1	0.1576	22.3045	0.0000	0.0071	0.1437	0.1714
rvrhon5	0.1496	8.3970	0.0000	0.0178	0.1147	0.1845
rvrhon20	0.4516	22.7346	0.0000	0.0199	0.4127	0.4906
rhrneg1	-1.2333	-11.0132	0.0000	0.1120	-1.4527	-1.0138
earn20	0.0816	7.3681	0.0000	0.0111	0.0599	0.1033
rhrneg5	-3.6200	-6.8504	0.0000	0.5284	-4.6557	-2.5843
rhrneg20	-7.4278	-6.3158	0.0000	1.1761	-9.7329	-5.1227

Table 6: EL-HAR: Pooled OLS results at $h = 20$

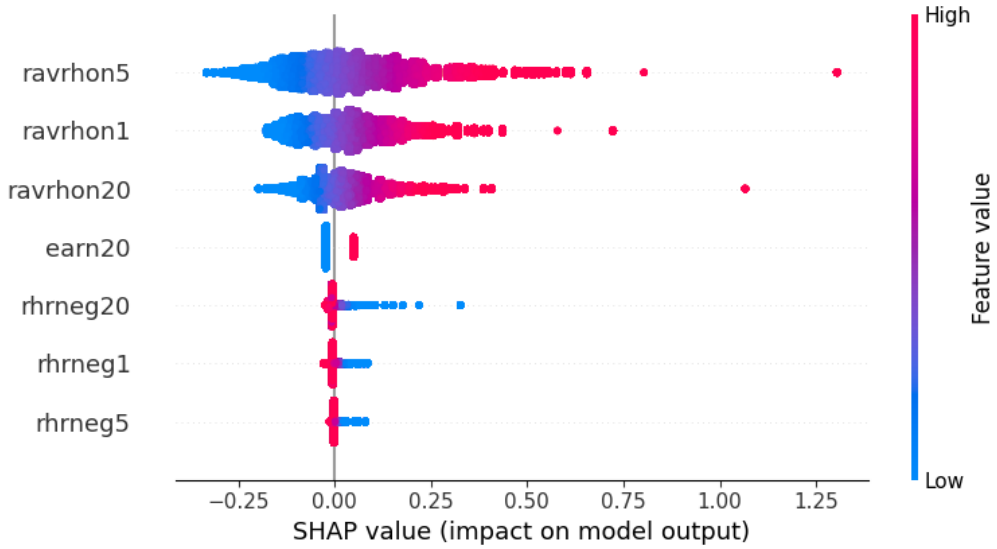

Figure 15: EL-HAR-RAV: LGBM SHAP values at $h = 20$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
ravrhon1	0.4757	12.0191	0.0000	0.0396	0.3982	0.5533
ravrhon20	2.1023	14.4481	0.0000	0.1455	1.8171	2.3875
rhrneg1	-0.9015	-7.8397	0.0000	0.1150	-1.1269	-0.6761
earn20	0.0837	7.5137	0.0000	0.0111	0.0619	0.1055
rhrneg5	-3.5427	-6.7701	0.0000	0.5233	-4.5684	-2.5171
ravrhon5	0.7143	5.2093	0.0000	0.1371	0.4455	0.9831
rhrneg20	-4.7032	-3.9848	0.0001	1.1803	-7.0166	-2.3899

Table 7: EL-HAR-RAV: Pooled OLS results at $h = 20$

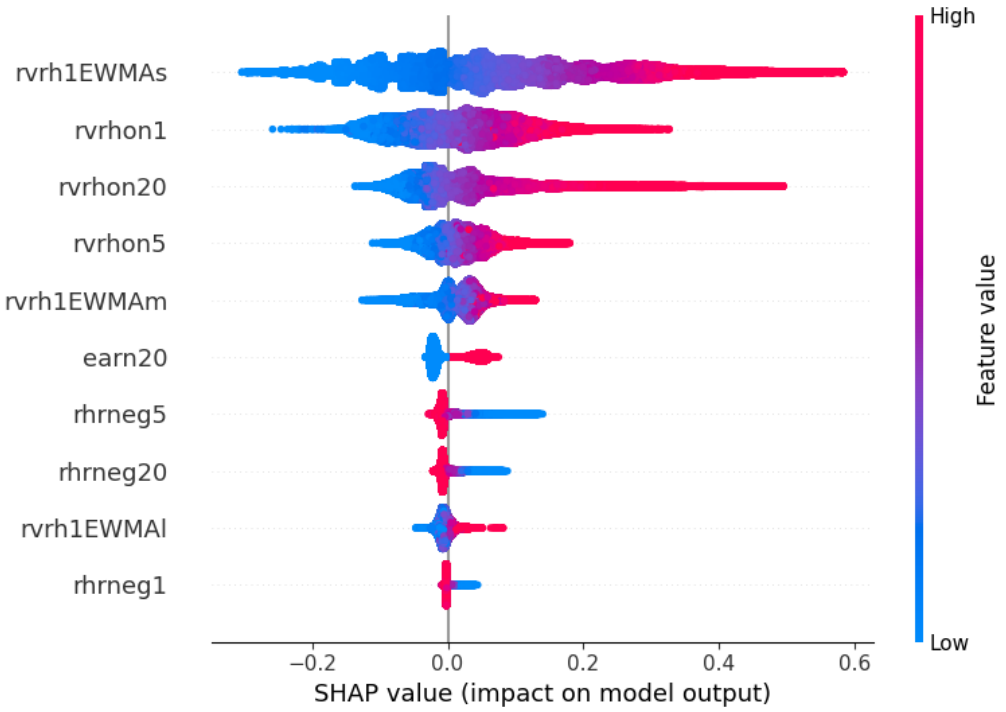

Figure 16: EL-HAR-ExpSML: LGBM SHAP values at $h = 20$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
rvrhon1	0.1561	21.9873	0.0000	0.0071	0.1422	0.1700
rvrhon5	0.1498	8.4170	0.0000	0.0178	0.1149	0.1847
rvrhon20	0.4308	20.9644	0.0000	0.0206	0.3906	0.4711
rhrneg1	-1.1617	-10.6285	0.0000	0.1093	-1.3759	-0.9475
earn20	0.0818	7.4139	0.0000	0.0110	0.0602	0.1034
rhrneg5	-3.4802	-6.4732	0.0000	0.5376	-4.5339	-2.4264
rhrneg20	-7.2997	-6.4032	0.0000	1.1400	-9.5340	-5.0653
rvrh1EWMAs	0.0308	2.3318	0.0197	0.0132	0.0049	0.0568
rvrh1EWMAl	0.0281	2.1690	0.0301	0.0129	0.0027	0.0534
rvrh1EWMAm	0.0178	0.8123	0.4166	0.0220	-0.0252	0.0609

Table 8: EL-HAR-ExpSML: Pooled OLS results at $h = 20$

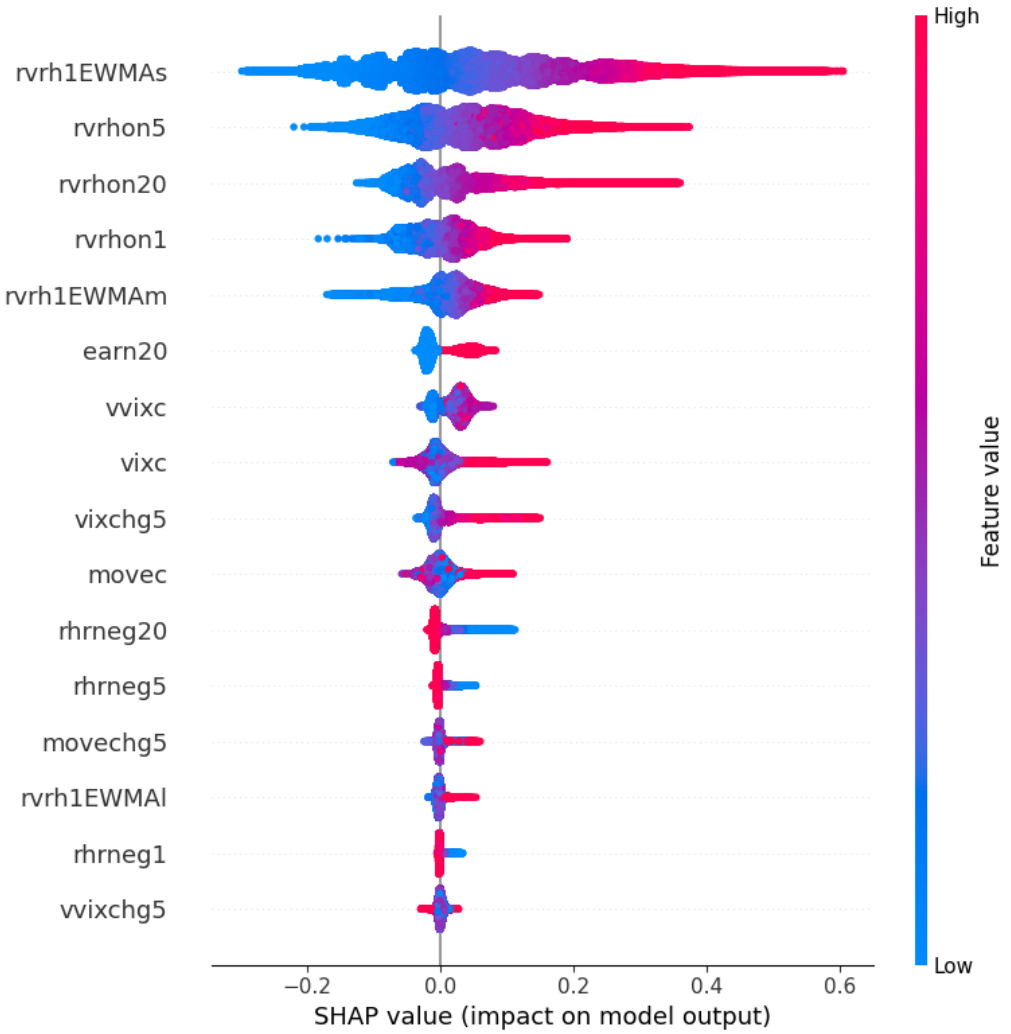

Figure 17: EL-HAR-ExpSML-RE: LGBM SHAP values at $h = 20$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
rvrhon1	0.1382	24.9220	0.0000	0.0055	0.1273	0.1490
rvrhon5	0.1455	8.9636	0.0000	0.0162	0.1136	0.1773
rvrhon20	0.4067	24.1425	0.0000	0.0168	0.3737	0.4398
rhrneg1	-0.9454	-8.9540	0.0000	0.1056	-1.1523	-0.7385
earn20	0.0846	8.0021	0.0000	0.0106	0.0639	0.1053
rhrneg20	-7.5771	-6.1194	0.0000	1.2382	-10.0040	-5.1503
rhrneg5	-2.4787	-5.1599	0.0000	0.4804	-3.4202	-1.5372
rvrh1EWMAl	0.0600	5.0437	0.0000	0.0119	0.0367	0.0834
rvrh1EWMAs	0.0434	3.3721	0.0007	0.0129	0.0182	0.0686
movec	0.0015	3.1325	0.0017	0.0005	0.0005	0.0024
vixchg5	0.1907	3.0838	0.0020	0.0618	0.0695	0.3119
movechg5	-0.1157	-1.6376	0.1015	0.0707	-0.2543	0.0228
vvixc	-0.0011	-1.5035	0.1327	0.0007	-0.0025	0.0003
rvrh1EWMAm	-0.0191	-0.8988	0.3687	0.0212	-0.0606	0.0225
vvixchg5	-0.0332	-0.4263	0.6699	0.0779	-0.1860	0.1195
vixc	-0.0063	-0.0444	0.9646	0.1426	-0.2859	0.2732

Table 9: EL-HAR-ExpSML-RE: Pooled OLS results at $h = 20$

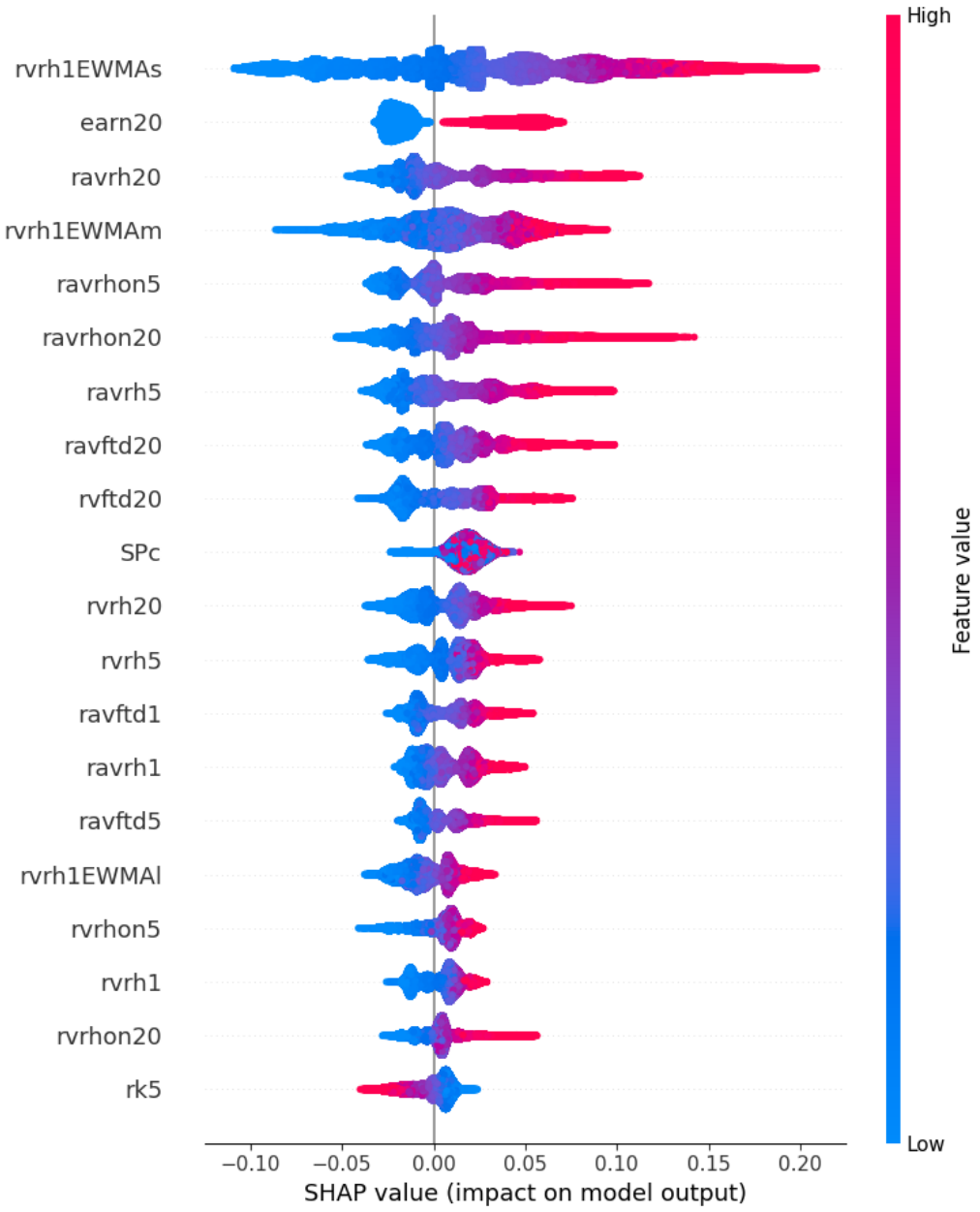

Figure 18: Model X: LGBM SHAP values at $h = 20$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
rvrhon1	0.1297	8.8358	0.0000	0.0147	0.1009	0.1585
rk5	-0.0119	-8.9875	0.0000	0.0013	-0.0144	-0.0093
rvrhon20	0.3554	8.1800	0.0000	0.0435	0.2703	0.4406
earn20	0.0821	7.5243	0.0000	0.0109	0.0607	0.1035
rvrhon5	0.3000	6.9754	0.0000	0.0430	0.2157	0.3843
rvftd1	0.0748	5.1536	0.0000	0.0145	0.0463	0.1032
rvrv5	-0.0173	-5.1001	0.0000	0.0034	-0.0239	-0.0106
rvrv20	-0.0088	-4.2793	0.0000	0.0021	-0.0129	-0.0048
rk20	-0.0103	-4.2591	0.0000	0.0024	-0.0150	-0.0055
rs5	-0.0153	-3.6601	0.0003	0.0042	-0.0235	-0.0071
vprh	-0.0000	-3.3628	0.0008	0.0000	-0.0000	-0.0000
rvrh1	-0.0947	-2.7293	0.0063	0.0347	-0.1627	-0.0267
rs20	0.0236	2.5689	0.0102	0.0092	0.0056	0.0416
rvnpm20	0.0288	2.2471	0.0246	0.0128	0.0037	0.0539
rvrh1EWMAl	0.0220	1.7788	0.0753	0.0124	-0.0022	0.0463
rvnpm5	-0.0191	-1.5647	0.1177	0.0122	-0.0429	0.0048
vvixc	-0.0016	-1.5402	0.1235	0.0010	-0.0036	0.0004
rvrh1EWMAm	0.0358	1.5343	0.1250	0.0233	-0.0099	0.0815
rvnpm1	-0.0070	-1.4542	0.1459	0.0048	-0.0163	0.0024
movec	0.0007	1.2688	0.2045	0.0005	-0.0004	0.0018

Table 10: Model X: Pooled OLS results at $h = 20$

8 Pseudo-Out-of-Sample Results

In this section, we evaluate the performance (accuracy) of each model on the pseudo-OOS test set. Table (11) compares the performance of OLS and LGBM using only the L-HAR features, as detailed in (38). LGBM has an OLS/LGBM ratio of around 1.08-1.10 for RMSPE and 1.06-1.08 for MAPE across all three horizons, suggesting that the OLS error is approximately 8-10% (6-8%) higher for RMSPE (MAPE). Including EA, as shown in Table (12), results in lower errors for both OLS and LGBM compared to the L-HAR model, though the ratios remain largely the same. Thus, both models appear equally adept at handling EA features. The EL-HAR-RAV model, presented in Table (13), offers substantial improvements over the previous two models. We suspect this is due to absolute returns being less sensitive to jumps, which are quite prevalent in our test sample for individual stocks during the Covid-19 crisis. The OLS/LGBM is ratio nearing 1, indicating that OLS has seen a significant relative improvement. Further, the

EL-HAR-ExpSML model presented in Table (14) performs similarly to EL-HAR, slightly better for OLS across all horizons, but worse for LGBM at the 1- and 5-day horizon, with some improvement at the 20-day horizon. Upon adding the RE features, as presented in Table (15), minor changes occur for LGBM while OLS has a fairly large reduction, especially at the medium and long horizon. Our final model, Model X, presented in Table (16), shows a marked improvement over the benchmark models. Nonetheless, OLS and LGBM perform on a more comparable level, with a ratio close to 1 at the 1- and 5-day horizon. LGBM performs significantly better at the 20-day horizon, with a ratio above 1.05 for both measures. Comparing the OLS benchmark OLS-BM using L-HAR with Model X using LGBM, we see ratios of 1.12-1.20, indicating a significant reduction.

L-HAR

	OLS-BM		LGBM-BM		OLS-BM/LGBM-BM	
h	RMSPE	MAPE	RMSPE	MAPE	RMSPE	MAPE
1	33.1070	25.3887	30.7413	24.0130	1.0770	1.0573
5	27.7009	21.1544	25.2968	19.5862	1.0950	1.0801
20	28.0788	21.6773	25.7795	20.0281	1.0892	1.0823

Table 11: LHS: mean results using pooled OLS. Middle: mean results using LGBM. RHS: OLS over LGBM. "BM" means benchmark.

EL-HAR

	OLS		LGBM		OLS/LGBM		OLS-BM/LGBM	
h	RMSPE	MAPE	RMSPE	MAPE	RMSPE	MAPE	RMSPE	MAPE
1	33.0772	25.3536	30.7204	23.9882	1.0767	1.0569	1.0777	1.0584
5	27.5154	20.9461	25.0722	19.3972	1.0974	1.0799	1.1048	1.0906
20	27.6339	21.1882	25.4200	19.5966	1.0871	1.0812	1.1046	1.1062

Table 12: LHS: mean results using pooled OLS and LGBM. RHS: OLS over LGBM and benchmark OLS over LGBM. Extended with "earn", "1bef", "2bef", "1aft", "2aft" for $h = 1$, "earn5" for $h = 5$, and "earn20" for $h = 20$.

EL-HAR-RAV

	OLS		LGBM		OLS/LGBM		OLS-BM/LGBM	
h	RMSPE	MAPE	RMSPE	MAPE	RMSPE	MAPE	RMSPE	MAPE
1	30.7606	24.1381	29.4323	23.1116	1.0451	1.0444	1.1249	1.0985
5	25.1526	19.7218	24.0675	18.6719	1.0451	1.0562	1.1510	1.1330
20	25.9314	20.2748	25.0184	19.1621	1.0365	1.0581	1.1223	1.1313

Table 13: LHS: mean results using pooled OLS and LGBM. RHS: OLS over LGBM and benchmark OLS over LGBM. Extended with "earn", "1bef", "2bef", "1aft", "2aft" for $h = 1$, "earn5" for $h = 5$, and "earn20" for $h = 20$. The "RAV" means that the RAV features have replaced the RV features for all horizons.

EL-HAR-ExpSML

	OLS		LGBM		OLS/LGBM		OLS-BM/LGBM	
h	RMSPE	MAPE	RMSPE	MAPE	RMSPE	MAPE	RMSPE	MAPE
1	32.9705	25.2983	30.9682	24.2302	1.0647	1.0441	1.0691	1.0478
5	27.3292	20.8244	25.2705	19.5013	1.0815	1.0678	1.0962	1.0848
20	27.2611	20.9220	24.9711	19.1345	1.0917	1.0934	1.1245	1.1329

Table 14: LHS: mean results using pooled OLS and LGBM. RHS: OLS over LGBM and benchmark OLS over LGBM. Extended with "earn", "1bef", "2bef", "1aft", "2aft" for $h = 1$, "earn5" for $h = 5$, and "earn20" for $h = 20$. The "ExpSML" means it includes expanding mean at the short, medium, and long-term horizons.

EL-HAR-ExpSML-RE

	OLS		LGBM		OLS/LGBM		OLS-BM/LGBM	
h	RMSPE	MAPE	RMSPE	MAPE	RMSPE	MAPE	RMSPE	MAPE
1	32.3823	25.0247	30.7742	24.0199	1.0523	1.0418	1.0758	1.0570
5	26.1004	20.0762	25.0420	19.4908	1.0423	1.0300	1.1062	1.0854
20	26.0749	20.2284	25.5606	20.2295	1.0201	0.9999	1.0985	1.0716

Table 15: LHS: mean results using pooled OLS and LGBM. RHS: OLS over LGBM and benchmark OLS over LGBM. Extended with "earn", "1bef", "2bef", "1aft", "2aft" for $h = 1$, "earn5" for $h = 5$, and "earn20" for $h = 20$. The "ExpSML" means it includes expanding mean at the short, medium, and long-term horizons, while the "RE" ("risk everywhere") means it includes closing values of VIX, VVIX, and MOVE, as well as 5-day percentage changes in VIX, VVIX, and MOVE.

Model X

	OLS		LGBM		OLS/LGBM		OLS-BM/LGBM	
h	RMSPE	MAPE	RMSPE	MAPE	RMSPE	MAPE	RMSPE	MAPE
1	29.1688	22.9527	28.6941	22.6201	1.0165	1.0147	1.1538	1.1224
5	24.2692	18.8849	23.7802	18.4282	1.0206	1.0248	1.1649	1.1479
20	24.7981	19.3069	23.5183	18.1060	1.0544	1.0663	1.1939	1.1972

Table 16: LHS: mean results using pooled OLS and LGBM. RHS: OLS over LGBM and benchmark OLS over LGBM. Extended with "earn", "1bef", "2bef", "1aft", "2aft" for $h = 1$, "earn5" for $h = 5$, and "earn20" for $h = 20$, and includes all other variables mentioned in Section (4).

9 Conclusion

In summary, our research underscores the potential of EA and OR in enhancing RV forecasting. Our approach integrated the squared overnight log return into the RV definition, aiming for more practically applicable forecasts via a full-day RV measure. This was done within a panel modeling context, utilizing extensions of the linear HAR model and its GB counterparts. With a dataset comprising 478 stocks and 2,086,068 daily observations, we investigated the issue via in-sample analysis and compared the pseudo-OOS results of both linear HAR models and GB models. Additional feature engineering and selection were performed to augment the models for further comparison and robustness testing.

Our data analysis reveals that the empirical distribution of our proposed full-day RV measure differs significantly from the academic intraday RV measure. A larger portion of the distribution is located at higher RV values, indicating significant ON volatility. We observed heightened volatility a few days before the announcement, a major volatility spike on the day of EA, and gradually declining yet elevated volatility post-EA. Moreover, HV stocks seem to exhibit much larger volatility spikes than LV stocks. Pooled OLS regression results confirm that all included EA features are highly significant and have reasonable coefficients corresponding with the rest of the analysis. Our study also highlighted substantial individual differences in the response of stock volatility to EA. Importantly, we found that EA days feature a volatility spike from close to open (ON), with high volatility persisting throughout the trading day following the release.

Moving forward, we outlined all models tested in detail and described the underlying theory of GB before presenting all in-sample results and pseudo-OOS results. GB is able to increase forecasting performance by around 5-10% compared to

the pooled OLS model with slight variations depending on which features are included. Model X, our best-performing model, sees a 12-20% improvement for both OLS and LGBM depending on the horizon. LGBM has the greatest relative performance against OLS looking at the L-HAR and EL-HAR model. Interestingly, the inclusion of EA dummy variables improves overall forecasting performance moderately, but to a lower degree than, e.g., simply substituting RAV for RV. We suspect this to be the case as we measure the mean of the total mean error of the individual stocks, where most days do not have an announcement (typically four per year), thus the total error will not be reduced that much. It could be that the model error is much lower on the days of EA and approximately unaffected otherwise, which will only reduce the total error slightly. The error reduction from including EA features is greater for longer forecasting horizons.

From the six models tested, the consensus seems to be that LGBM consistently outperforms OLS by approximately the same margin, regardless of the included features, albeit slightly more for less complex models. OLS behaves reasonably well to all included features, potentially due to limited nonlinearities and noise. While LGBM may be able to extract some additional performance unreachable by OLS, transitioning from a simple to a more complex model may not offer any additional relative improvement. The most complex model, Model X, achieved considerably lower model error than, for instance, L-HAR, which held true for both OLS and LGBM. The other models displayed varied results across horizons and model types (OLS or LGBM), but the simple EL-HAR-RAV model seems promising for both OLS and LGBM. In other words, using absolute values rather than squared values yielded significant improvements.

Several questions remain unanswered. First and foremost, focusing on days of EA and the model error on those days will give a better understanding of the use of EA features for short-term forecasting. Further, it would be intriguing to ascertain the extent of nonlinearities existing between different features, such as through the use of partial dependence plots. Examining the effects of interactions would also present another compelling aspect to investigate. Finally, it could be interesting to look at the FTD measure of volatility. The RHON measure ignores the volatility in prices in extended hours if the ON return is 0, although there may have been considerable extended-hour volatility. The FTD measure will capture such extended hour volatility and may show empirical improvements compared to the RHON measure used.

To conclude, our original contribution to the literature is threefold: We incor-

porate EA and offer a detailed analysis; We compare the RV forecasting performance of LightGBM (LGBM) with the simple HAR model; We apply the first and second points using a large dataset of individual stocks in a panel modeling setting, using a slightly different RV measure that incorporates ON information, along with additional feature selection and engineering.

APPENDIX

GBRT Algorithm

GBRT Algorithm - Text:

1. Initially, a base model is trained using the data.
2. The residuals of this base model, which are the discrepancies between predicted and actual values, are computed.
3. A subsequent model is then trained to predict these residuals.
4. After calculating the residuals of this second model, a third model is trained to predict these residuals.
5. This process is repeated until either a predetermined number of models have been trained or a certain performance level has been reached.
6. The predictions of each individual model are then aggregated to provide a forecast for a new instance.
7. The contribution of each model to the final prediction is determined by the learning rate parameter. Although smaller learning rates may require more models to be trained, they could enhance the generalization performance of the model.

GBRT Algorithm - Mathematical Notation:

1. Initialize $f_0(x) = \arg\min_\gamma \sum_{i=1}^{N} L\left(y_i, \gamma\right)$.
2. For $m = 1$ to M:
   (a) For $i = 1, 2, \ldots, N$ compute
   $$r_{im} = -\left[\frac{\partial L\left(y_i, f\left(x_i\right)\right)}{\partial f\left(x_i\right)}\right]_{f=f_{m-1}}.$$
   (b) Fit a regression tree to the targets r_{im} giving terminal regions $R_{jm}, j = 1, 2, \ldots, J_m$.
   (c) For $j = 1, 2, \ldots, J_m$ compute
   $$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L\left(y_i, f_{m-1}\left(x_i\right) + \gamma\right).$$
   (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I\left(x \in R_{jm}\right)$.

3. Output $\hat{f}(x) = f_M(x)$.

For additional information, refer to Bentéjac et al. (2021) and Hastie et al. (2009).

Data Analysis

As illustrated in Figures (19) and (20), we observe both high and persistent autocorrelation. RV (RAV) starts with an autocorrelation of approximately 0.6 (0.8), which gradually decreases to around 0.2-0.3 (0.4) after 50 lags. While the different RV measures display slightly varied initial autocorrelations and persistence, the RAV measures follow similar patterns. As for the PACFs, depicted in Figures (21) and (22), they die off after around 10 lags. The Figures (23), (24), (25), and (26) show the ACF and PACF values of RV and RAV in NPM. Although the initial values are much lower, there appears to be persistence.

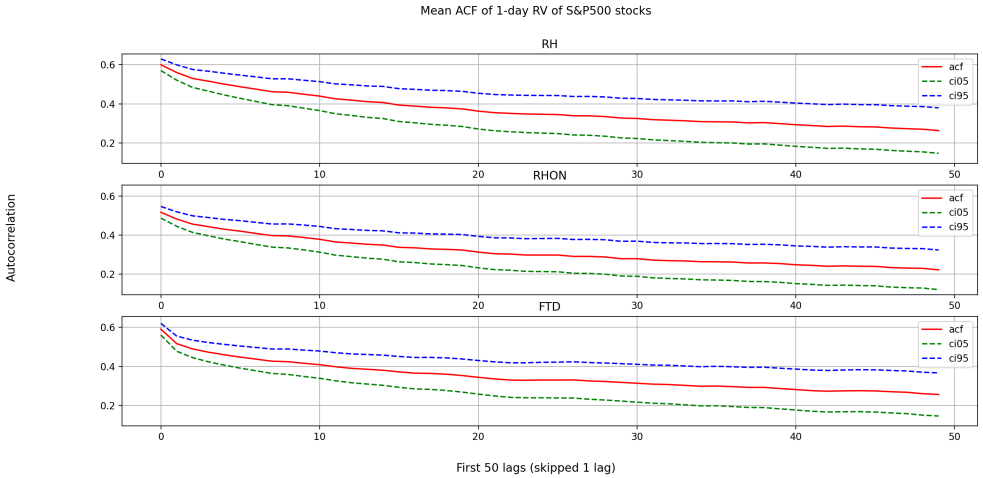

Figure 19: Mean ACF of 1-day RV of the S&P 500 stocks with 5% and 95% confidence intervals. Illustrated for the three different measures of volatility.

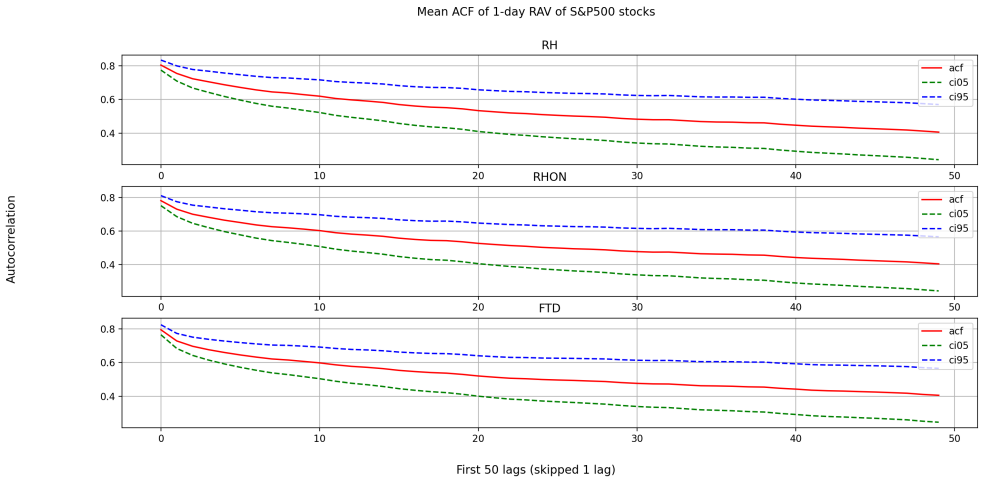

Figure 20: Mean ACF of 1-day RAV of the S&P 500 stocks with 5% and 95% confidence intervals. Illustrated for the three different measures of volatility.

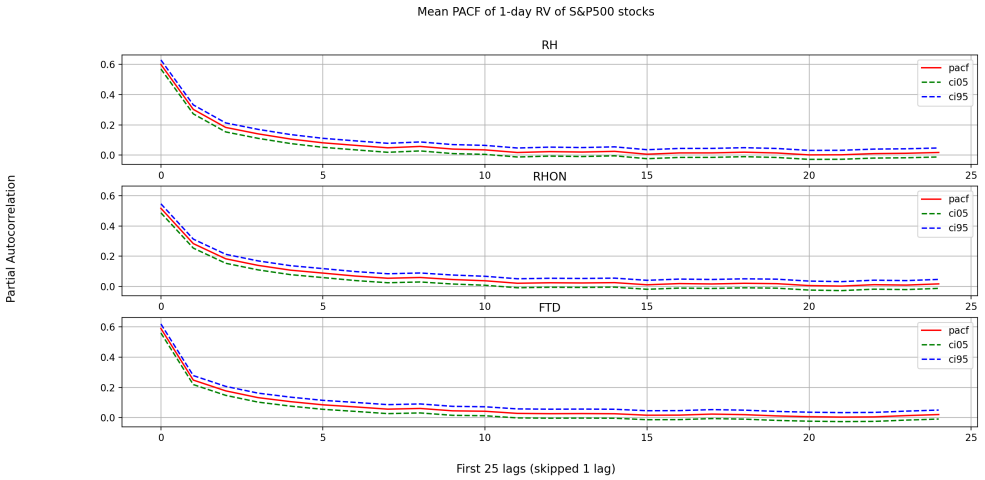

Figure 21: Mean PACF of 1-day RV of the S&P 500 stocks with 5% and 95% confidence intervals. Illustrated for the three different measures of volatility.

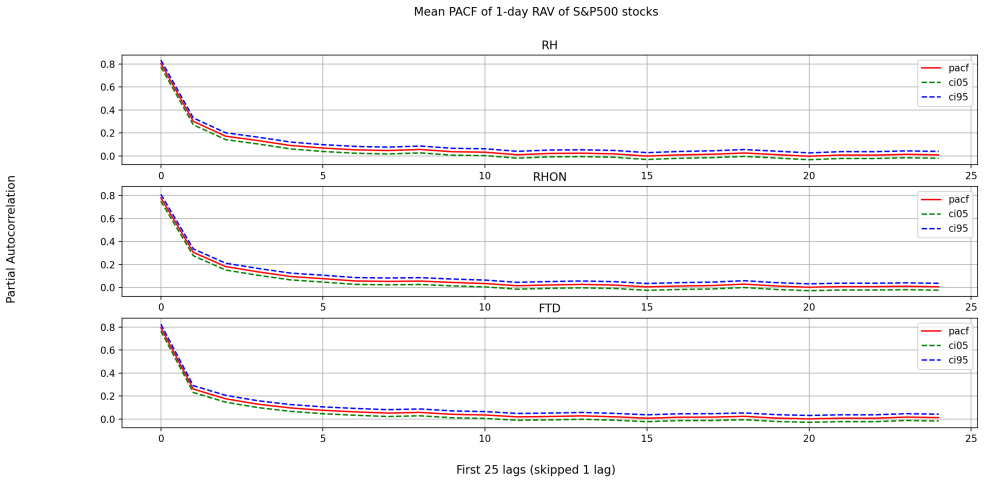

Figure 22: Mean PACF of 1-day RAV of the S&P 500 stocks with 5% and 95% confidence intervals. Illustrated for the three different measures of volatility.

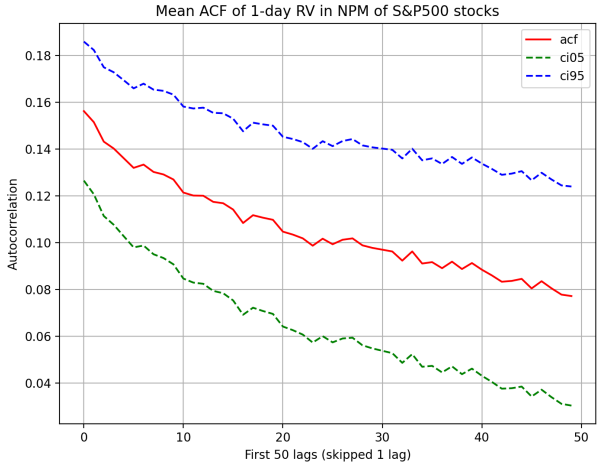

Figure 23: Mean ACF of 1-day RV in NPM of the S&P 500 stocks with 5% and 95% confidence intervals. Illustrated for the three different measures of volatility.

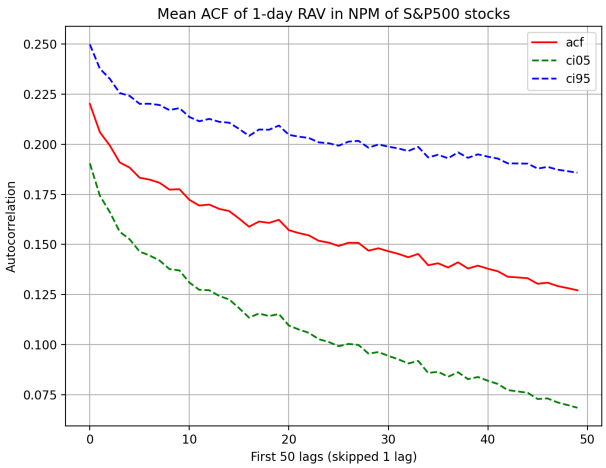

Figure 24: Mean ACF of 1-day RAV in NPM of the S&P 500 stocks with 5% and 95% confidence intervals. Illustrated for the three different measures of volatility.

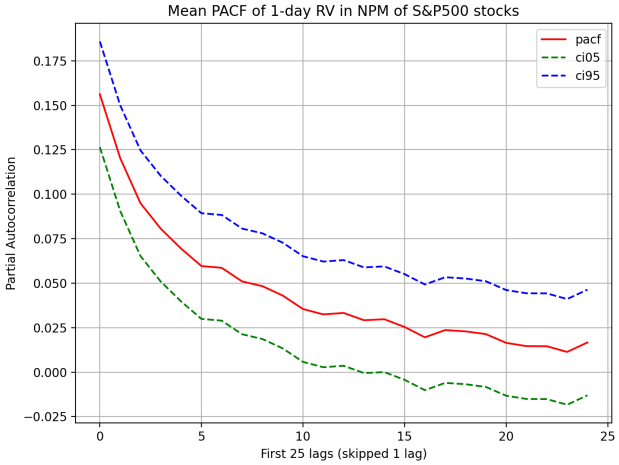

Figure 25: Mean PACF of 1-day RV in NPM of the S&P 500 stocks with 5% and 95% confidence intervals. Illustrated for the three different measures of volatility.

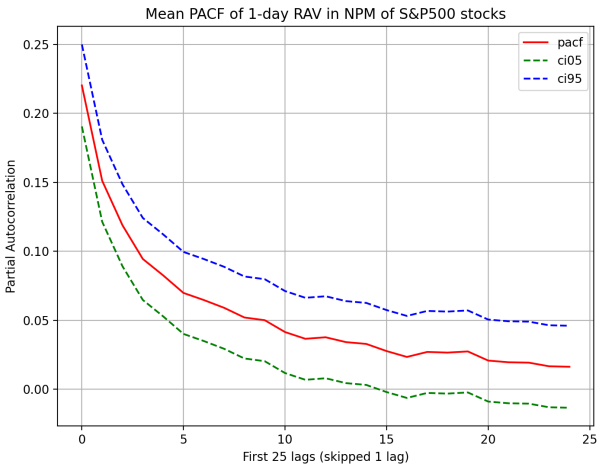

Figure 26: Mean PACF of 1-day RAV in NPM of the S&P 500 stocks with 5% and 95% confidence intervals. Illustrated for the three different measures of volatility.

Optimized Hyperparameters

L-HAR

h	LR	NOL	FF	MGTS	ET	L1	L2
1	0.0268	37	0.5671	3.4790	False	2.4755	0.5768
5	0.0976	56	0.6139	7.3811	False	2.0705	1.8768
20	0.0947	100	0.5605	0.5216	False	2.9736	1.1286

EL-HAR

h	LR	NOL	FF	MGTS	ET	L1	L2
1	0.0203	25	0.6838	2.2439	False	2.5049	2.7925
5	0.0318	27	0.6637	5.7496	False	1.7602	1.5126
20	0.0807	42	0.5958	5.2029	False	0.5517	2.4881

EL-HAR-RAV

h	LR	NOL	FF	MGTS	ET	L1	L2
1	0.0947	38	0.2349	2.1811	False	0.2014	1.5318
5	0.0080	30	0.3899	6.9888	False	2.1324	1.2498
20	0.0929	85	0.1985	2.5075	False	2.7267	0.8100

EL-HAR-ExpSML

h	LR	NOL	FF	MGTS	ET	L1	L2
1	0.0333	37	0.5620	7.8879	False	1.2262	0.9031
5	0.0821	78	0.3890	0.0341	False	1.0592	2.5254
20	0.0496	60	0.3614	1.4624	False	1.4770	0.6262

EL-HAR-ExpSML-RE

h	LR	NOL	FF	MGTS	ET	L1	L2
1	0.0422	23	0.9346	5.4343	True	6.6332	2.6914
5	0.0116	11	0.3293	0.6725	False	6.7039	4.1805
20	0.0533	96	0.4268	7.1572	False	7.7923	6.5040

Model X

h	LR	NOL	FF	MGTS	ET	L1	L2
1	0.0052	93	0.1162	4.8425	True	0.0500	9.5052
5	0.0775	24	0.5198	3.1218	True	4.3597	4.0184
20	0.0198	163	0.2542	8.6695	True	8.6043	3.9358

Table 17: Hyperparameters for all LGBM models. Abbreviations: LR: learning rate; NOL: number of leaves; FF: feature fraction; MGTS: minimum gain to split; ET: extra trees; L1: (lambda) L1 regularization; L2: (lambda) L2 regularization.

SHAP Values and OLS Results

L-HAR

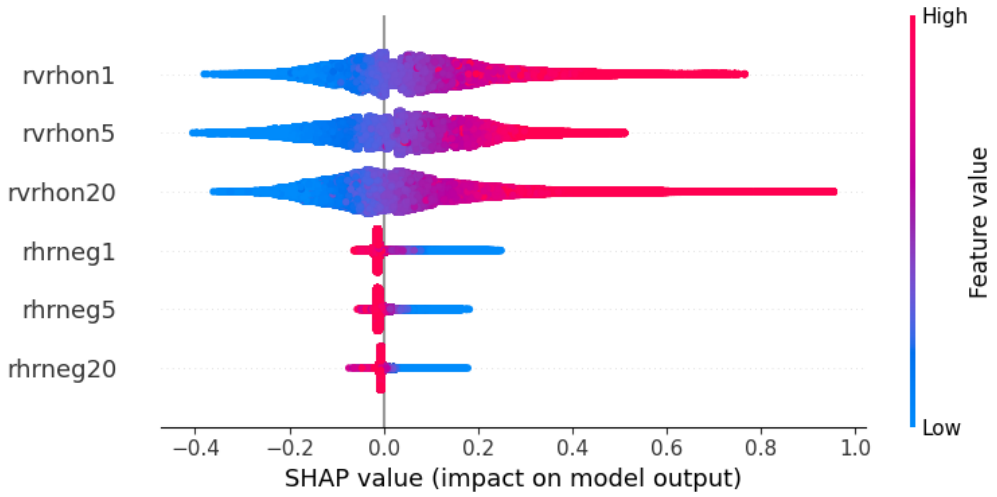

Figure 27: L-HAR: LGBM SHAP values at $h = 1$

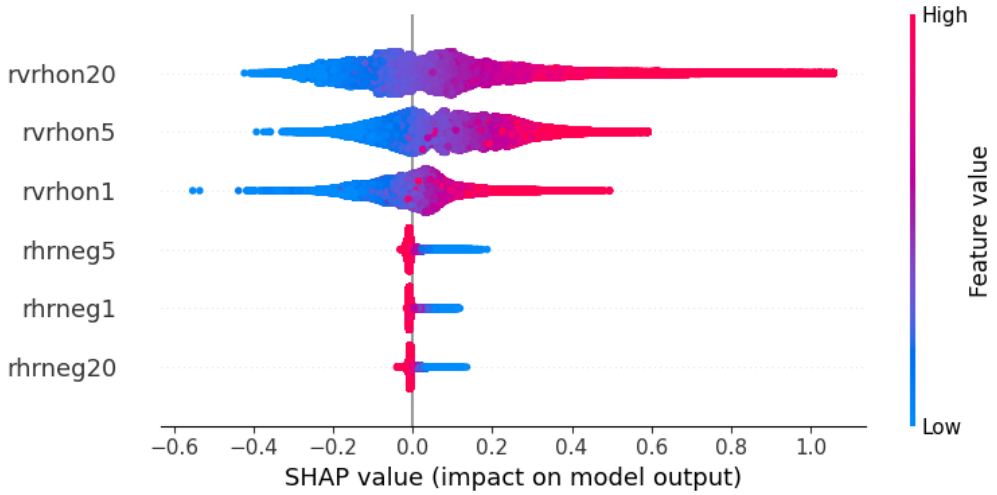

Figure 28: L-HAR: LGBM SHAP values at $h = 5$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
rvrhon1	0.2745	32.4660	0.0000	0.0085	0.2579	0.2910
rvrhon5	0.2286	20.5882	0.0000	0.0111	0.2069	0.2504
rvrhon20	0.3489	28.9699	0.0000	0.0120	0.3253	0.3725
rhrneg1	-2.3493	-13.0623	0.0000	0.1799	-2.7018	-1.9968
rhrneg20	-5.6524	-7.0603	0.0000	0.8006	-7.2216	-4.0833
rhrneg5	-3.5975	-6.9961	0.0000	0.5142	-4.6053	-2.5896

Table 18: L-HAR: Pooled OLS results at $h = 1$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
rvrhon1	0.2126	32.8708	0.0000	0.0065	0.1999	0.2252
rvrhon5	0.1985	14.5501	0.0000	0.0136	0.1717	0.2252
rvrhon20	0.4036	26.6716	0.0000	0.0151	0.3740	0.4333
rhrneg1	-1.7169	-13.6173	0.0000	0.1261	-1.9640	-1.4698
rhrneg20	-7.6408	-7.8503	0.0000	0.9733	-9.5485	-5.7332
rhrneg5	-3.3081	-5.6421	0.0000	0.5863	-4.4573	-2.1589

Table 19: L-HAR: Pooled OLS results at $h = 5$

EL-HAR

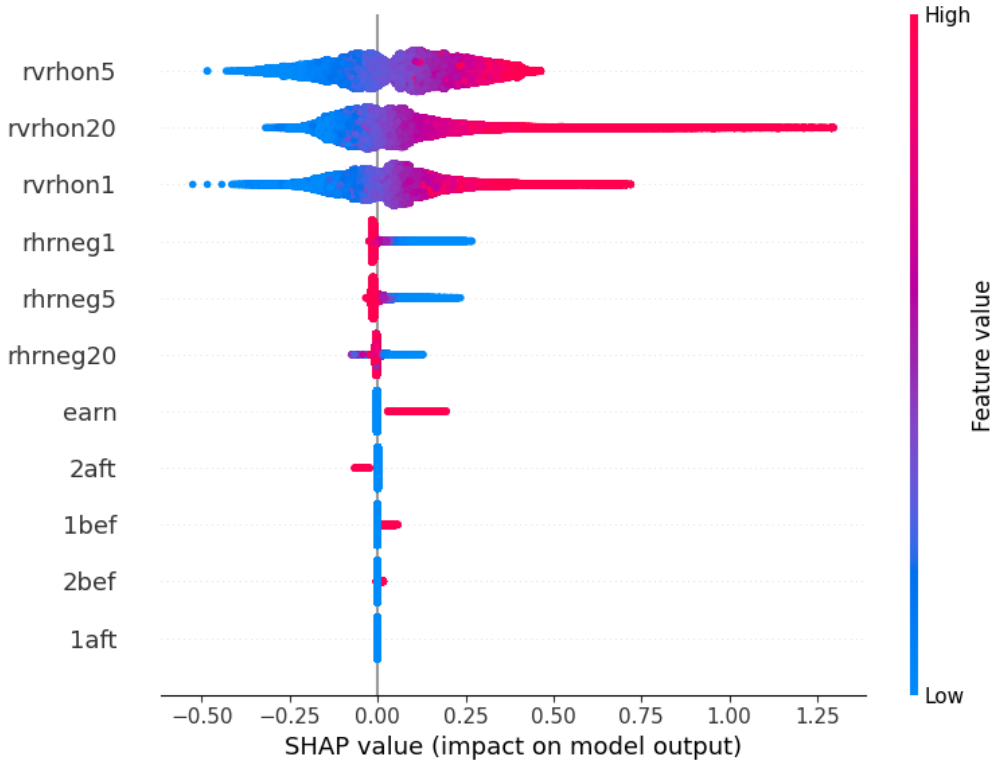

Figure 29: EL-HAR: LGBM SHAP values at $h = 1$

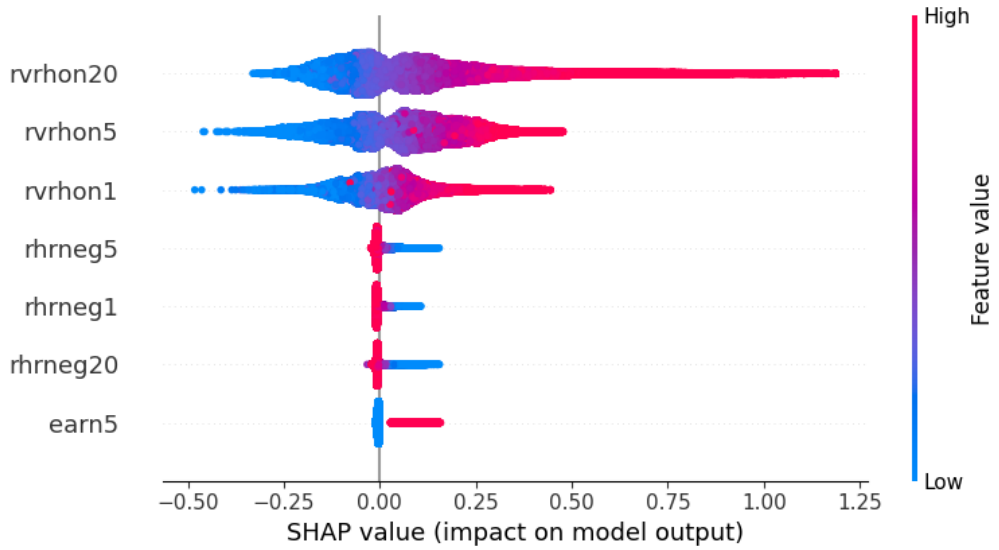

Figure 30: EL-HAR: LGBM SHAP values at $h = 5$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
rvrhon1	0.2737	32.4607	0.0000	0.0084	0.2572	0.2903
rvrhon5	0.2294	20.7175	0.0000	0.0111	0.2077	0.2511
rvrhon20	0.3485	29.0432	0.0000	0.0120	0.3250	0.3721
rhrneg1	-2.3449	-13.0578	0.0000	0.1796	-2.6969	-1.9930
earn	0.1643	13.1056	0.0000	0.0125	0.1398	0.1889
2aft	-0.0917	-10.6532	0.0000	0.0086	-0.1085	-0.0748
rhrneg20	-5.6741	-7.0815	0.0000	0.8012	-7.2445	-4.1036
rhrneg5	-3.6195	-7.0481	0.0000	0.5135	-4.6260	-2.6130
1bef	0.0556	6.4909	0.0000	0.0086	0.0388	0.0724
1aft	-0.0235	-2.8513	0.0044	0.0082	-0.0397	-0.0074
2bef	0.0171	2.2509	0.0244	0.0076	0.0022	0.0320

Table 20: EL-HAR: Pooled OLS results at $h = 1$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
rvrhon1	0.2129	32.9801	0.0000	0.0065	0.2003	0.2256
rvrhon5	0.1937	14.0921	0.0000	0.0137	0.1668	0.2207
rvrhon20	0.4056	26.8864	0.0000	0.0151	0.3760	0.4351
rhrneg1	-1.7070	-13.6527	0.0000	0.1250	-1.9520	-1.4619
earn5	0.1086	9.7913	0.0000	0.0111	0.0868	0.1303
rhrneg20	-7.8236	-8.0166	0.0000	0.9759	-9.7364	-5.9108
rhrneg5	-3.3537	-5.6989	0.0000	0.5885	-4.5071	-2.2003

Table 21: EL-HAR: Pooled OLS results at $h = 5$

EL-HAR-RAV

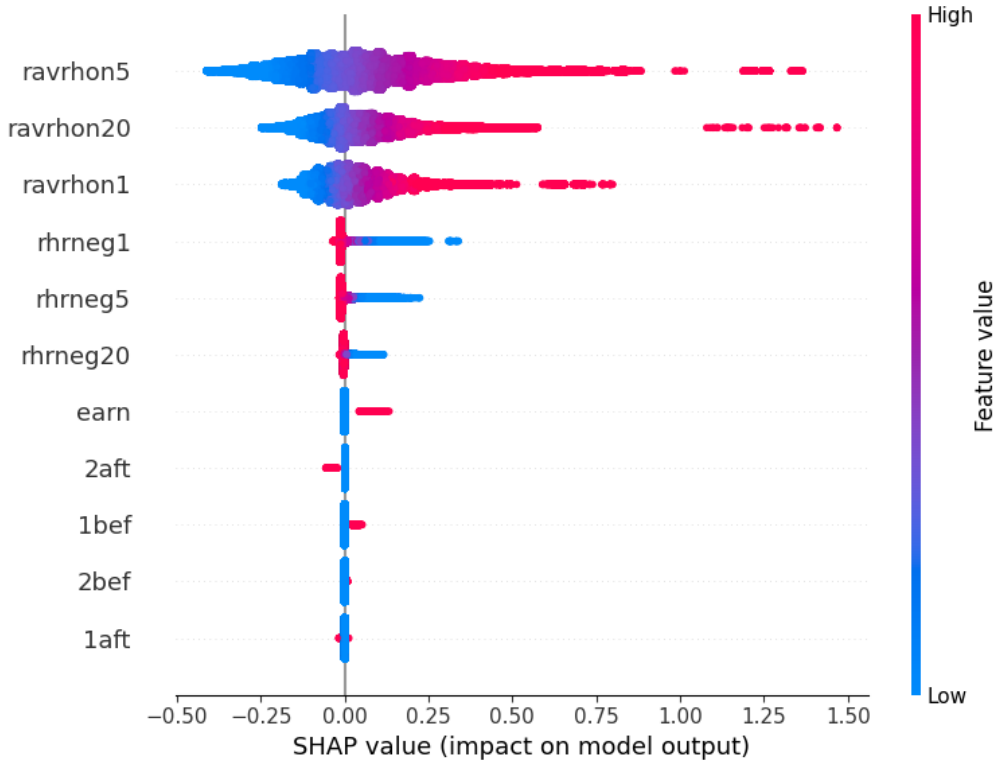

Figure 31: EL-HAR-RAV: LGBM SHAP values at $h = 1$

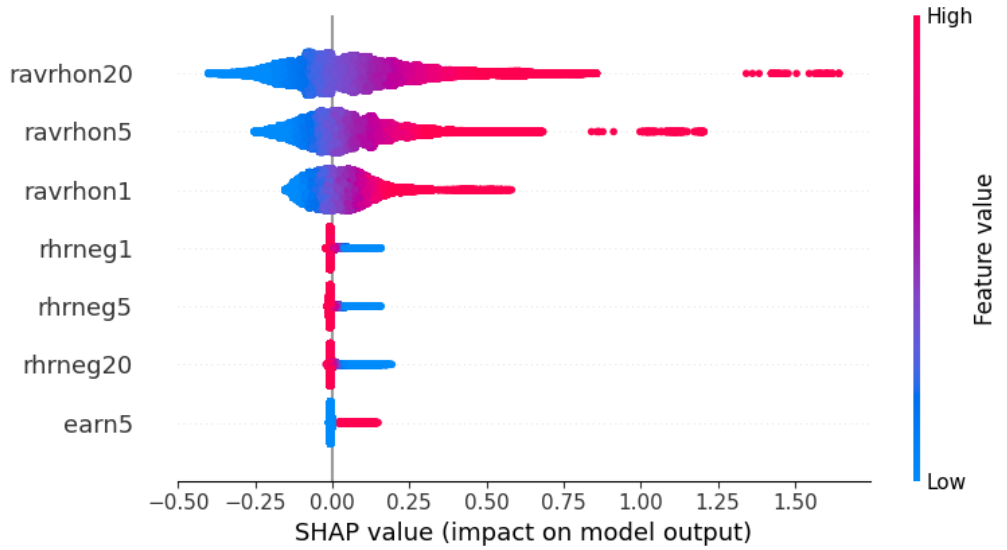

Figure 32: EL-HAR-RAV: LGBM SHAP values at $h = 5$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
ravrhon1	1.2955	27.9516	0.0000	0.0463	1.2046	1.3863
ravrhon5	1.1186	17.4266	0.0000	0.0642	0.9928	1.2444
ravrhon20	1.2375	16.9460	0.0000	0.0730	1.0943	1.3806
rhrneg1	-1.7060	-10.0520	0.0000	0.1697	-2.0386	-1.3733
earn	0.1788	14.6019	0.0000	0.0122	0.1548	0.2028
1bef	0.0677	8.3268	0.0000	0.0081	0.0518	0.0837
2aft	-0.0584	-7.7935	0.0000	0.0075	-0.0731	-0.0437
rhrneg5	-2.8320	-6.3357	0.0000	0.4470	-3.7081	-1.9559
2bef	0.0298	4.1323	0.0000	0.0072	0.0157	0.0440
rhrneg20	-2.6724	-4.0127	0.0001	0.6660	-3.9777	-1.3671
1aft	0.0112	1.5142	0.1300	0.0074	-0.0033	0.0256

Table 22: EL-HAR-RAV: Pooled OLS results at $h = 1$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
ravrhon1	0.8350	22.4794	0.0000	0.0371	0.7622	0.9078
ravrhon5	0.9913	10.3000	0.0000	0.0962	0.8027	1.1799
ravrhon20	1.6815	16.1119	0.0000	0.1044	1.4770	1.8861
rhrneg1	-1.2442	-10.2247	0.0000	0.1217	-1.4827	-1.0057
earn5	0.1227	10.9348	0.0000	0.0112	0.1007	0.1446
rhrneg20	-4.8256	-5.2096	0.0000	0.9263	-6.6411	-3.0101
rhrneg5	-2.9049	-5.1744	0.0000	0.5614	-4.0053	-1.8046

Table 23: EL-HAR-RAV: Pooled OLS results at $h = 5$

EL-HAR-ExpSML

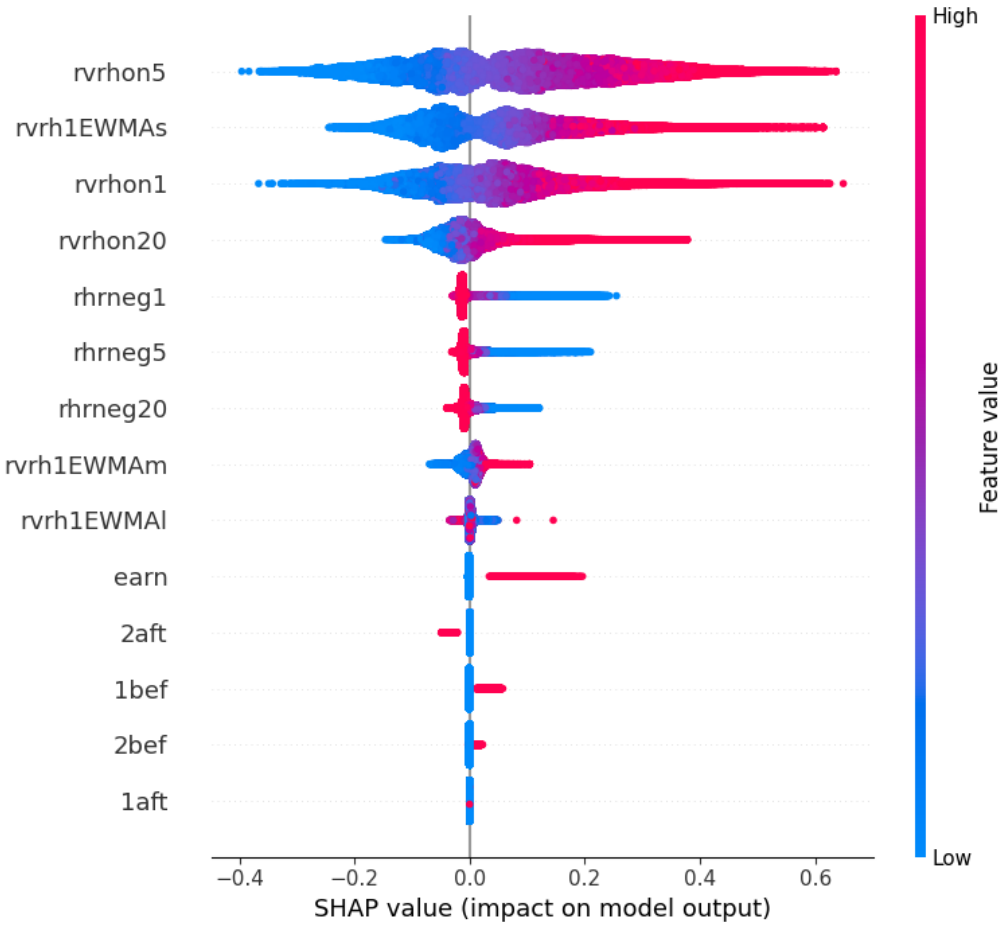

Figure 33: EL-HAR-ExpSML: LGBM SHAP values at $h = 1$

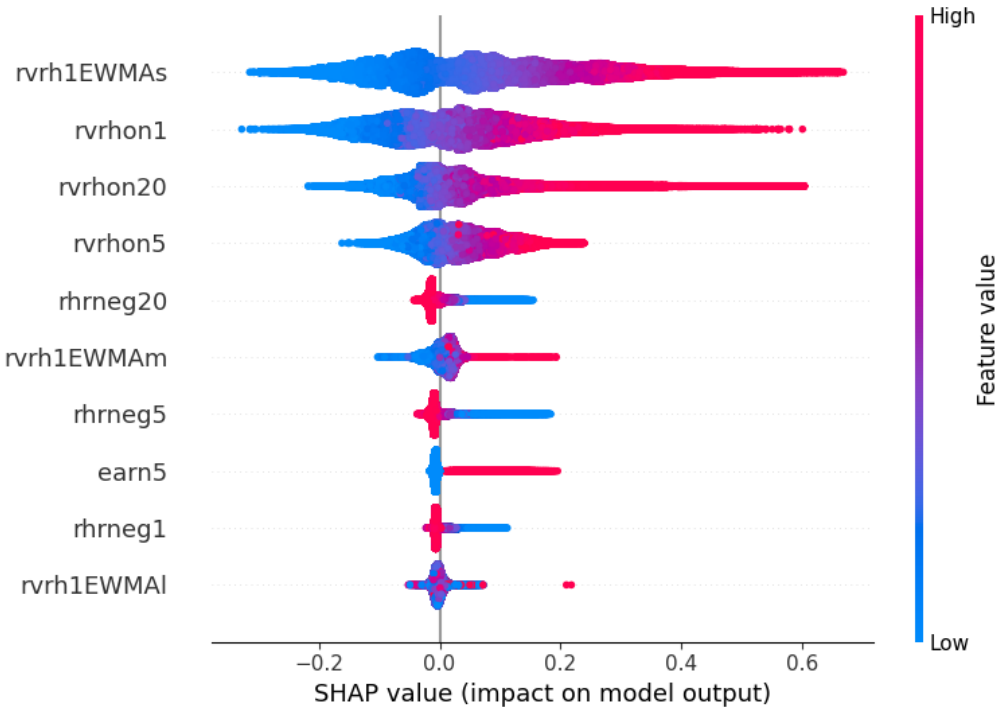

Figure 34: EL-HAR-ExpSML: LGBM SHAP values at $h = 5$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
rvrhon1	0.2730	32.2374	0.0000	0.0085	0.2564	0.2896
rvrhon5	0.2295	20.6975	0.0000	0.0111	0.2078	0.2512
rvrhon20	0.3391	27.7384	0.0000	0.0122	0.3151	0.3630
rhrneg1	-2.3148	-12.8124	0.0000	0.1807	-2.6689	-1.9607
earn	0.1652	13.1440	0.0000	0.0126	0.1405	0.1898
2aft	-0.0904	-10.5007	0.0000	0.0086	-0.1073	-0.0736
rhrneg20	-5.6324	-7.1095	0.0000	0.7922	-7.1852	-4.0797
rhrneg5	-3.5525	-6.8500	0.0000	0.5186	-4.5690	-2.5360
1bef	0.0562	6.5839	0.0000	0.0085	0.0395	0.0730
1aft	-0.0223	-2.6883	0.0072	0.0083	-0.0385	-0.0060
rvrh1EWMAm	0.0282	2.5882	0.0096	0.0109	0.0068	0.0496
2bef	0.0177	2.3318	0.0197	0.0076	0.0028	0.0326
rvrh1EWMAs	0.0057	0.7906	0.4292	0.0073	-0.0085	0.0200
rvrh1EWMAl	0.0014	0.2151	0.8297	0.0065	-0.0113	0.0141

Table 24: EL-HAR-ExpSML: Pooled OLS results at $h = 1$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
rvrhon1	0.2119	32.4643	0.0000	0.0065	0.1991	0.2247
rvrhon5	0.1938	14.0404	0.0000	0.0138	0.1668	0.2209
rvrhon20	0.3915	25.1866	0.0000	0.0155	0.3610	0.4219
rhrneg1	-1.6603	-13.3852	0.0000	0.1240	-1.9034	-1.4172
earn5	0.1094	9.8694	0.0000	0.0111	0.0877	0.1311
rhrneg20	-7.7488	-8.0591	0.0000	0.9615	-9.6334	-5.8643
rhrneg5	-3.2574	-5.4896	0.0000	0.5934	-4.4204	-2.0944
rvrh1EWMAs	0.0160	1.6584	0.0972	0.0097	-0.0029	0.0349
rvrh1EWMAl	0.0139	1.5484	0.1215	0.0090	-0.0037	0.0315
rvrh1EWMAm	0.0229	1.5012	0.1333	0.0152	-0.0070	0.0527

Table 25: EL-HAR-ExpSML: Pooled OLS results at $h = 5$

EL-HAR-ExpSML-RE

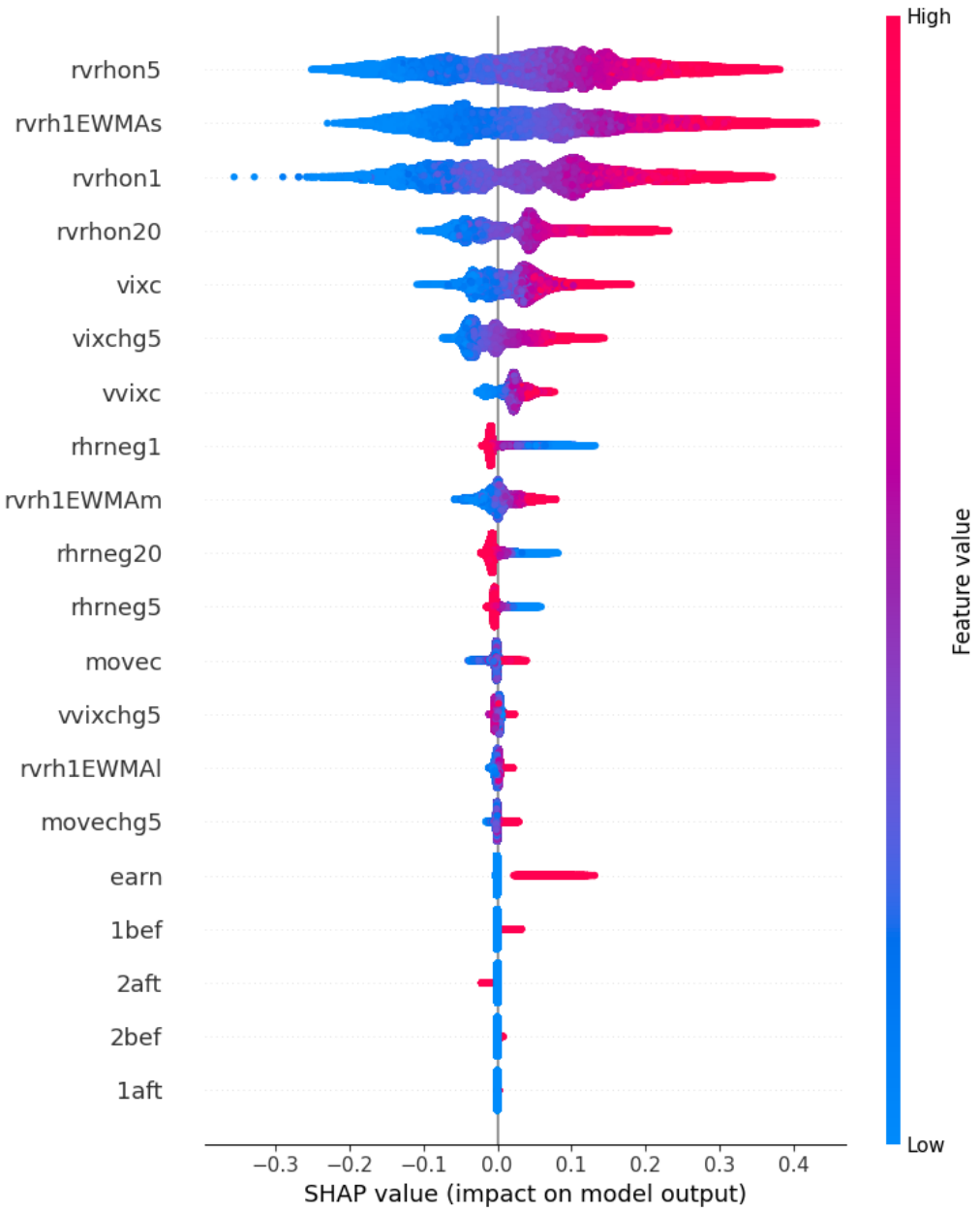

Figure 35: EL-HAR-ExpSML-RE: LGBM SHAP values at $h = 1$

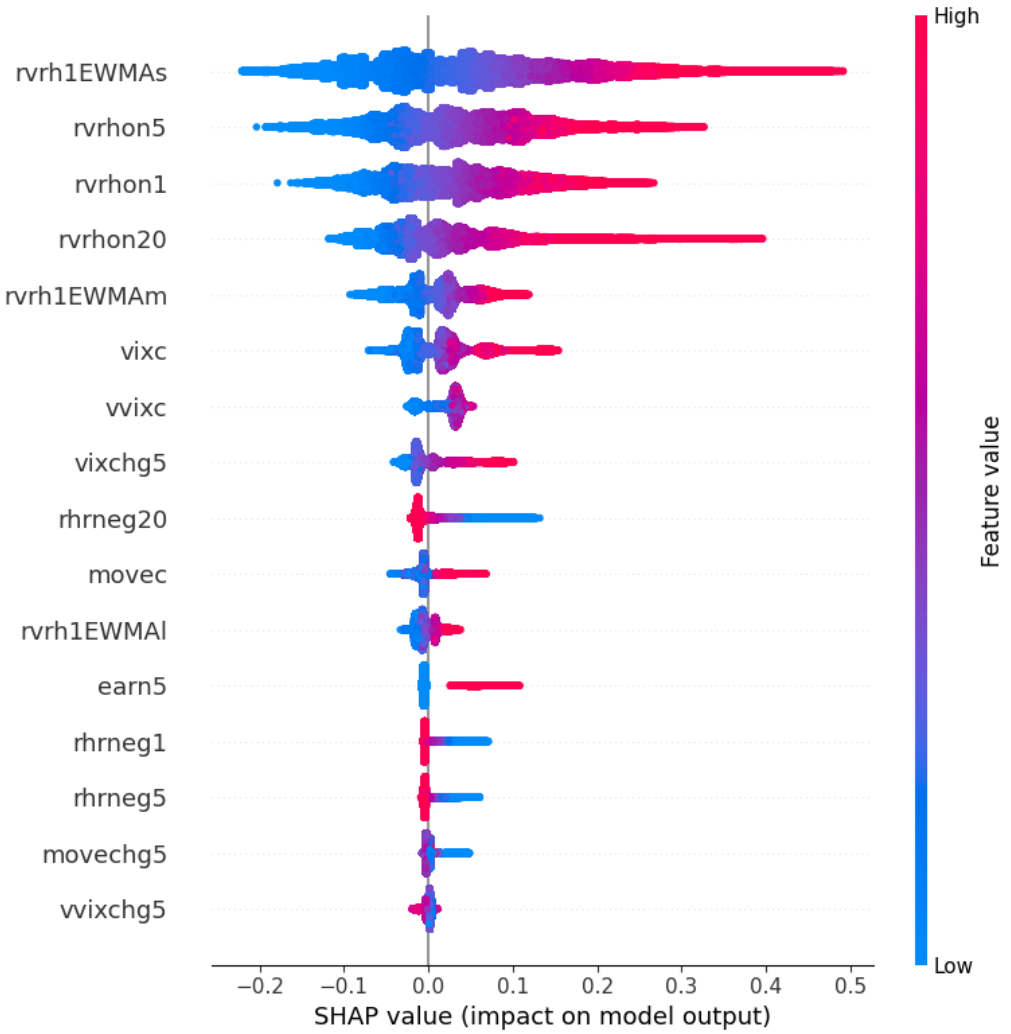

Figure 36: EL-HAR-ExpSML-RE: LGBM SHAP values at $h = 5$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
rvrhon1	0.2339	32.2909	0.0000	0.0072	0.2197	0.2481
earn	0.1715	14.0034	0.0000	0.0122	0.1475	0.1955
rhrneg1	-1.8848	-10.2955	0.0000	0.1831	-2.2436	-1.5260
2aft	-0.0820	-12.3027	0.0000	0.0067	-0.0951	-0.0690
rvrhon5	0.2115	18.4695	0.0000	0.0115	0.1891	0.2340
rvrhon20	0.3264	32.5329	0.0000	0.0100	0.3067	0.3461
vixc	0.5922	8.2621	0.0000	0.0717	0.4517	0.7327
1bef	0.0610	7.8303	0.0000	0.0078	0.0458	0.0763
vixchg5	0.2904	7.5733	0.0000	0.0383	0.2153	0.3656
rhrneg20	-4.8688	-4.8129	0.0000	1.0116	-6.8515	-2.8861
rvrh1EWMAl	0.0281	4.5431	0.0000	0.0062	0.0160	0.0403
2bef	0.0231	3.4068	0.0007	0.0068	0.0098	0.0364
rhrneg5	-1.4720	-3.0584	0.0022	0.4813	-2.4154	-0.5287
rvrh1EWMAs	0.0184	2.2900	0.0220	0.0081	0.0027	0.0342
vvixchg5	-0.0807	-1.6087	0.1077	0.0501	-0.1789	0.0176
1aft	-0.0087	-1.2773	0.2015	0.0068	-0.0220	0.0046
movec	0.0002	1.2143	0.2246	0.0002	-0.0001	0.0006
rvrh1EWMAm	0.0055	0.4820	0.6298	0.0115	-0.0170	0.0281
vvixc	-0.0000	-0.1308	0.8959	0.0004	-0.0007	0.0007
movechg5	0.0021	0.0552	0.9560	0.0376	-0.0716	0.0757

Table 26: EL-HAR-ExpSML-RE: Pooled OLS results at $h = 1$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
rvrhon1	0.1821	35.3842	0.0000	0.0051	0.1720	0.1922
rvrhon5	0.1823	13.8983	0.0000	0.0131	0.1566	0.2081
rvrhon20	0.3717	28.2744	0.0000	0.0131	0.3459	0.3974
rhrneg1	-1.3030	-10.2779	0.0000	0.1268	-1.5515	-1.0545
earn5	0.1140	10.6350	0.0000	0.0107	0.0930	0.1350
rhrneg20	-7.3551	-6.3755	0.0000	1.1536	-9.6162	-5.0940
vixchg5	0.2448	4.9661	0.0000	0.0493	0.1482	0.3413
rvrh1EWMAl	0.0422	4.9334	0.0000	0.0086	0.0254	0.0590
vixc	0.3944	4.1453	0.0000	0.0951	0.2079	0.5808
rhrneg5	-1.6532	-3.2116	0.0013	0.5148	-2.6621	-0.6443
rvrh1EWMAs	0.0285	2.8831	0.0039	0.0099	0.0091	0.0478
movec	0.0006	2.1339	0.0329	0.0003	0.0001	0.0012
movechg5	-0.0845	-1.7564	0.0790	0.0481	-0.1788	0.0098
vvixchg5	-0.0583	-0.9212	0.3569	0.0632	-0.1822	0.0657
vvixc	-0.0005	-0.8726	0.3829	0.0005	-0.0015	0.0006
rvrh1EWMAm	-0.0049	-0.3171	0.7512	0.0153	-0.0349	0.0252

Table 27: EL-HAR-ExpSML-RE: Pooled OLS results at $h = 5$

Model X

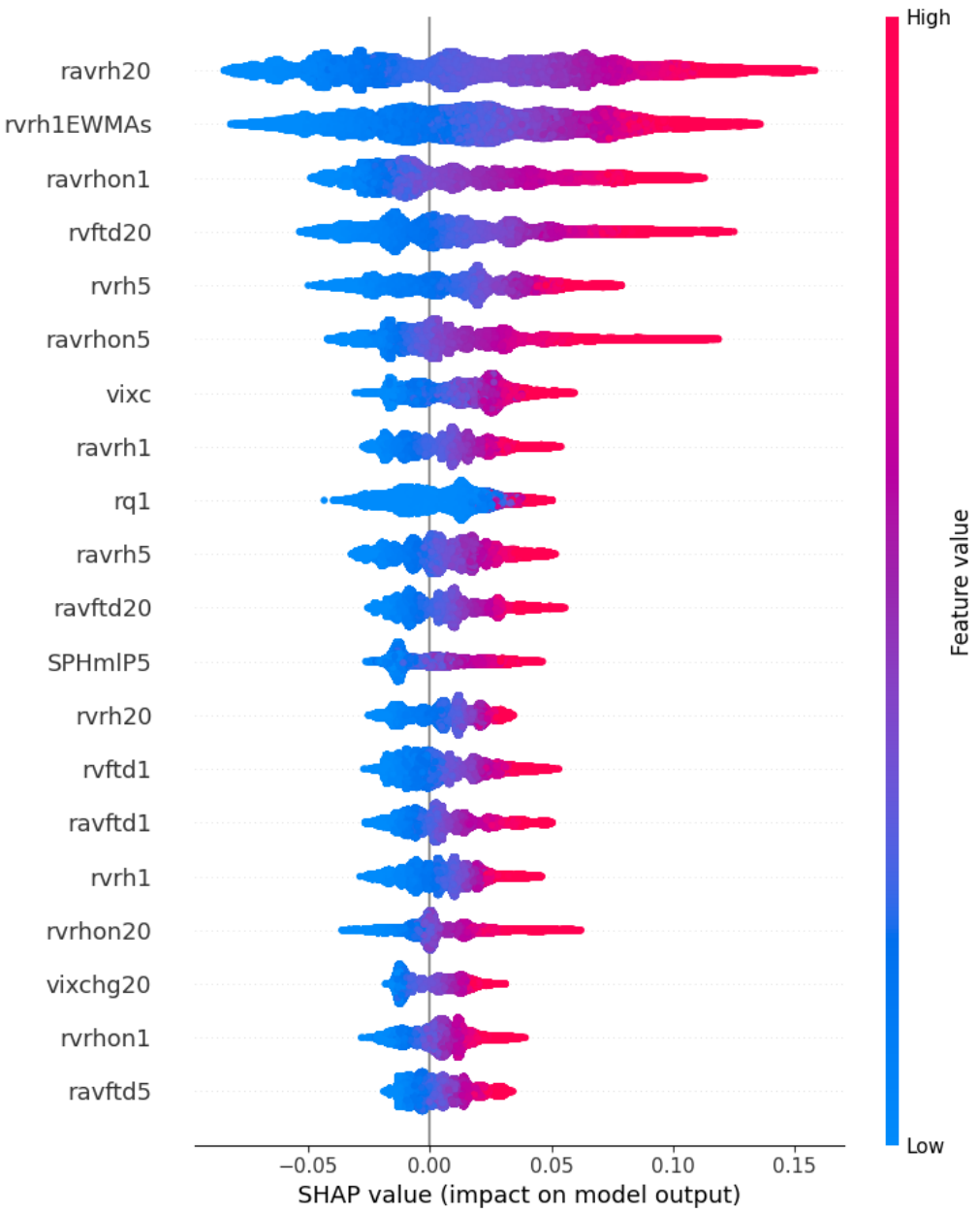

Figure 37: Model X: LGBM SHAP values at $h = 1$

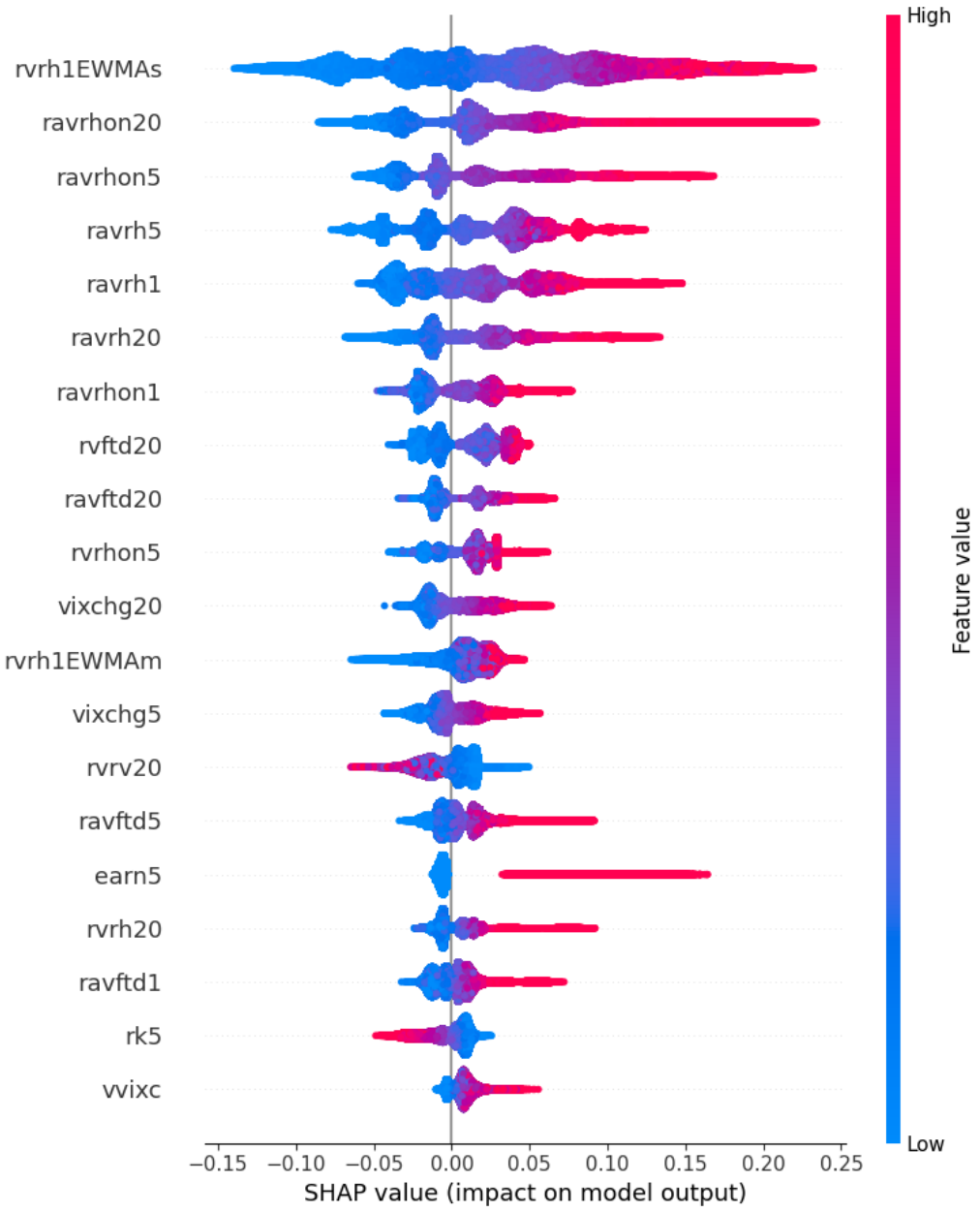

Figure 38: Model X: LGBM SHAP values at $h = 5$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
rvrhon5	0.3918	10.7180	0.0000	0.0366	0.3202	0.4635
rk5	-0.0166	-15.0597	0.0000	0.0011	-0.0188	-0.0145
rvrhon1	0.2197	10.3538	0.0000	0.0212	0.1781	0.2613
rvrhon20	0.2623	9.0325	0.0000	0.0290	0.2054	0.3193
vprh	0.0000	9.7742	0.0000	0.0000	0.0000	0.0000
rvftd1	0.2372	6.3488	0.0000	0.0374	0.1639	0.3104
rvrv20	-0.0091	-5.7537	0.0000	0.0016	-0.0122	-0.0060
rs5	-0.0169	-4.9473	0.0000	0.0034	-0.0237	-0.0102
rvrv5	-0.0165	-4.9444	0.0000	0.0033	-0.0231	-0.0100
rvrh1	-0.2112	-4.6490	0.0000	0.0454	-0.3002	-0.1221
SPc	-0.0000	-3.8145	0.0001	0.0000	-0.0000	-0.0000
vixchg20	0.0980	3.0990	0.0019	0.0316	0.0360	0.1599
rvrh1EWMAm	0.0303	2.2043	0.0275	0.0137	0.0034	0.0572
rvnpm20	0.0227	2.0428	0.0411	0.0111	0.0009	0.0444
rk20	-0.0028	-2.0140	0.0440	0.0014	-0.0055	-0.0001
rvnpm5	-0.0200	-1.8434	0.0653	0.0108	-0.0412	0.0013
rvnpm1	-0.0102	-1.6700	0.0949	0.0061	-0.0222	0.0018
rvrh5	-0.0712	-1.4007	0.1613	0.0508	-0.1708	0.0284
rvftd5	0.0435	1.3284	0.1840	0.0327	-0.0207	0.1077
rvftd20	0.0395	1.0437	0.2966	0.0379	-0.0347	0.1137

Table 28: Model X: Pooled OLS results at $h = 1$

	Coef.	t-stat	p-val	Std. Err.	Low. CI.	Upp. CI.
earn5	0.1169	10.6473	0.0000	0.0110	0.0954	0.1384
rvrhon5	0.3524	9.0201	0.0000	0.0391	0.2758	0.4290
rvrhon20	0.3156	8.6060	0.0000	0.0367	0.2437	0.3875
rvrhon1	0.1696	10.7528	0.0000	0.0158	0.1387	0.2005
rk5	-0.0154	-13.9622	0.0000	0.0011	-0.0175	-0.0132
rvrv20	-0.0100	-5.5211	0.0000	0.0018	-0.0136	-0.0065
rvftd1	0.1273	5.4672	0.0000	0.0233	0.0817	0.1729
rvrv5	-0.0174	-5.1057	0.0000	0.0034	-0.0241	-0.0107
rs5	-0.0139	-3.6916	0.0002	0.0038	-0.0214	-0.0065
rvrh1	-0.1328	-3.4344	0.0006	0.0387	-0.2086	-0.0570
rk20	-0.0056	-3.0945	0.0020	0.0018	-0.0091	-0.0020
rvnpm20	0.0281	2.3343	0.0196	0.0121	0.0045	0.0518
vixchg20	0.0926	2.2181	0.0265	0.0418	0.0108	0.1745
rvrh1EWMAm	0.0320	1.7047	0.0882	0.0188	-0.0048	0.0688
SPc	-0.0000	-1.6663	0.0956	0.0000	-0.0000	0.0000
rvnpm1	-0.0082	-1.6349	0.1021	0.0050	-0.0180	0.0016
rvnpm5	-0.0193	-1.5937	0.1110	0.0121	-0.0431	0.0044
rvrh1EWMAl	0.0125	1.4010	0.1612	0.0089	-0.0050	0.0301
rvftd5	0.0365	1.2031	0.2289	0.0304	-0.0230	0.0960
movec	0.0003	0.9602	0.3370	0.0004	-0.0004	0.0011

Table 29: Model X: Pooled OLS results at $h = 5$

Data Description

Ticker	Start	End	Total days	Years	# Trading days
A	2005-01-03	2022-10-21	6,500	17.80	4,483
AAP	2005-01-03	2022-10-21	6,500	17.80	4,483
AAPL	2005-01-03	2022-10-21	6,500	17.80	4,483
ABC	2005-01-03	2022-10-21	6,500	17.80	4,483
ABMD	2005-01-03	2022-10-21	6,500	17.80	4,481
ABT	2005-01-03	2022-10-21	6,500	17.80	4,483
ADBE	2005-01-03	2022-10-21	6,500	17.80	4,483
ADI	2005-01-03	2022-10-21	6,500	17.80	4,483
ADM	2005-01-03	2022-10-21	6,500	17.80	4,483
ADP	2005-01-03	2022-10-21	6,500	17.80	4,483
ADSK	2005-01-03	2022-10-21	6,500	17.80	4,483
AEE	2005-01-03	2022-10-21	6,500	17.80	4,483
AEP	2005-01-03	2022-10-21	6,500	17.80	4,483
AES	2005-01-03	2022-10-21	6,500	17.80	4,483
AFL	2005-01-03	2022-10-21	6,500	17.80	4,483
AIG	2005-01-03	2022-10-21	6,500	17.80	4,483
AIZ	2005-01-03	2022-10-21	6,500	17.80	4,483
AJG	2005-01-03	2022-10-21	6,500	17.80	4,483
AKAM	2005-01-03	2022-10-21	6,500	17.80	4,483
ALB	2005-01-03	2022-10-21	6,500	17.80	4,483
ALGN	2007-05-01	2022-10-21	5,652	15.47	3,899
ALK	2005-01-03	2022-10-21	6,500	17.80	4,483
ALL	2005-01-03	2022-10-21	6,500	17.80	4,483
AMAT	2005-01-03	2022-10-21	6,500	17.80	4,483
AMD	2005-01-03	2022-10-21	6,500	17.80	4,483

AME	2005-01-03	2022-10-21	6,500	17.80	4,483
AMG	2005-01-03	2022-10-21	6,500	17.80	4,483
AMGN	2005-01-03	2022-10-21	6,500	17.80	4,483
AMP	2005-10-03	2022-10-21	6,227	17.05	4,294
AMZN	2005-01-03	2022-10-21	6,500	17.80	4,483
AN	2005-01-03	2022-10-21	6,500	17.80	4,483
ANF	2005-01-03	2022-10-21	6,500	17.80	4,483
ANSS	2005-01-03	2022-10-21	6,500	17.80	4,483
AON	2007-04-30	2022-10-21	5,653	15.48	3,899
AOS	2007-04-02	2022-10-21	5,681	15.55	3,919
APA	2005-01-03	2022-10-21	6,500	17.80	4,483
APD	2005-01-03	2022-10-21	6,500	17.80	4,483
APH	2005-01-03	2022-10-21	6,500	17.80	4,483
ARE	2005-01-03	2022-10-21	6,500	17.80	4,483
ASH	2005-01-03	2022-10-21	6,500	17.80	4,483
ATGE	2007-05-01	2022-10-21	5,652	15.47	3,898
ATI	2005-01-03	2022-10-21	6,500	17.80	4,483
ATO	2005-01-03	2022-10-21	6,500	17.80	4,483
ATVI	2007-04-30	2022-10-21	5,653	15.48	3,880
AVB	2005-01-03	2022-10-21	6,500	17.80	4,483
AVY	2005-01-03	2022-10-21	6,500	17.80	4,483
AXP	2005-01-03	2022-10-21	6,500	17.80	4,483
AYI	2005-01-03	2022-10-21	6,500	17.80	4,483
AZO	2005-01-03	2022-10-21	6,500	17.80	4,483
BA	2005-01-03	2022-10-21	6,500	17.80	4,483
BAC	2005-01-03	2022-10-21	6,500	17.80	4,483
BALL	2005-01-03	2022-10-21	6,500	17.80	4,483
BAX	2005-01-03	2022-10-21	6,500	17.80	4,483
BBBY	2005-01-03	2022-10-21	6,500	17.80	4,483
BBWI	2007-04-27	2022-10-21	5,656	15.49	3,900
BBY	2005-01-03	2022-10-21	6,500	17.80	4,483
BC	2005-01-03	2022-10-21	6,500	17.80	4,483
BDX	2005-01-03	2022-10-21	6,500	17.80	4,483
BEN	2005-01-03	2022-10-21	6,500	17.80	4,483
BIG	2006-09-01	2022-10-21	5,894	16.14	4,063
BIIB	2005-01-03	2022-10-21	6,500	17.80	4,482
BIO	2007-05-08	2022-10-21	5,645	15.46	3,894
BK	2007-07-02	2022-10-21	5,590	15.30	3,856
BKNG	2007-04-30	2022-10-21	5,653	15.48	3,899
BLK	2005-01-03	2022-10-21	6,500	17.80	4,483
BMY	2005-01-03	2022-10-21	6,500	17.80	4,483
BR	2007-05-01	2022-10-21	5,652	15.47	3,899
BRO	2007-01-03	2022-10-21	5,770	15.80	3,980
BSX	2005-01-03	2022-10-21	6,500	17.80	4,483
BWA	2005-01-03	2022-10-21	6,500	17.80	4,483
BXP	2005-01-03	2022-10-21	6,500	17.80	4,483
CAG	2005-01-03	2022-10-21	6,500	17.80	4,483
CAH	2005-01-03	2022-10-21	6,500	17.80	4,483
CAR	2006-09-05	2022-10-21	5,890	16.13	4,062
CAT	2005-01-03	2022-10-21	6,500	17.80	4,483
CCI	2005-01-03	2022-10-21	6,500	17.80	4,483
CCK	2005-01-03	2022-10-21	6,500	17.80	4,483
CCL	2005-01-03	2022-10-21	6,500	17.80	4,483
CDNS	2005-10-31	2022-10-21	6,199	16.97	4,274
CE	2005-03-01	2022-10-21	6,443	17.64	4,444
CF	2005-10-03	2022-10-21	6,227	17.05	4,294
CHD	2004-01-02	2022-10-21	6,867	18.80	4,735
CHRW	2006-01-03	2022-10-21	6,135	16.80	4,231
CI	2004-01-02	2022-10-21	6,867	18.80	4,735
CIEN	2006-12-01	2022-10-21	5,803	15.89	4,000
CINF	2004-01-02	2022-10-21	6,867	18.80	4,735
CL	2004-01-02	2022-10-21	6,867	18.80	4,735
CLF	2004-01-02	2022-10-21	6,867	18.80	4,735
CLX	2004-01-02	2022-10-21	6,867	18.80	4,735
CMA	2004-01-02	2022-10-21	6,867	18.80	4,735
CMCSA	2004-01-02	2022-10-21	6,867	18.80	4,735
CME	2004-01-02	2022-10-21	6,867	18.80	4,735
CMG	2006-01-26	2022-10-21	6,112	16.73	4,215
CMI	2004-01-02	2022-10-21	6,867	18.80	4,735
CMS	2004-01-02	2022-10-21	6,867	18.80	4,735
CNC	2004-01-02	2022-10-21	6,867	18.80	4,735
CNP	2004-01-02	2022-10-21	6,867	18.80	4,735
CNX	2004-01-02	2022-10-21	6,867	18.80	4,735
COF	2004-01-02	2022-10-21	6,867	18.80	4,735
COO	2004-01-02	2022-10-21	6,867	18.80	4,735
COP	2004-01-02	2022-10-21	6,867	18.80	4,735
COST	2007-04-30	2022-10-21	5,653	15.48	3,900
CPB	2004-01-02	2022-10-21	6,867	18.80	4,735
CPRT	2004-01-02	2022-10-21	6,867	18.80	4,735
CPT	2007-01-03	2022-10-21	5,770	15.80	3,980
CRL	2007-01-03	2022-10-21	5,770	15.80	3,980

CRM	2004-06-23	2022-10-21	6,694	18.33	4,617
CSCO	2004-01-02	2022-10-21	6,867	18.80	4,735
CSX	2004-01-02	2022-10-21	6,867	18.80	4,735
CTAS	2004-01-02	2022-10-21	6,867	18.80	4,735
CTRA	2007-04-27	2022-10-21	5,656	15.49	3,878
CTSH	2004-01-02	2022-10-21	6,867	18.80	4,735
CVS	2004-01-02	2022-10-21	6,867	18.80	4,735
CVX	2004-01-02	2022-10-21	6,867	18.80	4,735
D	2005-01-03	2022-10-21	6,500	17.80	4,483
DAL	2007-05-16	2022-10-21	5,637	15.43	3,888
DDS	2005-01-03	2022-10-21	6,500	17.80	4,483
DE	2005-01-03	2022-10-21	6,500	17.80	4,483
DFS	2007-07-02	2022-10-21	5,590	15.30	3,856
DGX	2005-01-03	2022-10-21	6,500	17.80	4,483
DHI	2005-01-03	2022-10-21	6,500	17.80	4,483
DHR	2005-01-03	2022-10-21	6,500	17.80	4,483
DIS	2005-01-03	2022-10-21	6,500	17.80	4,483
DISH	2005-01-03	2022-10-21	6,500	17.80	4,483
DLR	2005-01-03	2022-10-21	6,500	17.80	4,483
DLTR	2007-04-27	2022-10-21	5,656	15.49	3,901
DLX	2005-01-03	2022-10-21	6,500	17.80	4,483
DOV	2005-01-03	2022-10-21	6,500	17.80	4,483
DPZ	2005-01-03	2022-10-21	6,500	17.80	4,483
DRE	2005-01-03	2022-09-30	6,479	17.74	4,468
DRI	2005-01-03	2022-10-21	6,500	17.80	4,483
DTE	2005-01-03	2022-10-21	6,500	17.80	4,483
DVA	2005-01-03	2022-10-21	6,500	17.80	4,483
DVN	2005-01-03	2022-10-21	6,500	17.80	4,483
DXCM	2005-05-02	2022-10-21	6,381	17.47	4,400
EA	2007-04-30	2022-10-21	5,653	15.48	3,899
EBAY	2005-01-03	2022-10-21	6,500	17.80	4,483
ECL	2005-01-03	2022-10-21	6,500	17.80	4,483
ED	2005-01-03	2022-10-21	6,500	17.80	4,483
EFX	2005-01-03	2022-10-21	6,500	17.80	4,483
EIX	2007-04-30	2022-10-21	5,653	15.48	3,900
EL	2005-01-03	2022-10-21	6,500	17.80	4,483
EMN	2005-01-03	2022-10-21	6,500	17.80	4,483
EMR	2005-01-03	2022-10-21	6,500	17.80	4,483
EOG	2005-01-03	2022-10-21	6,500	17.80	4,483
EQIX	2005-01-03	2022-10-21	6,500	17.80	4,483
EQR	2005-01-03	2022-10-21	6,500	17.80	4,483
EQT	2005-01-03	2022-10-21	6,500	17.80	4,483
ESS	2005-01-03	2022-10-21	6,500	17.80	4,483
ETR	2005-01-03	2022-10-21	6,500	17.80	4,483
EW	2005-01-03	2022-10-21	6,500	17.80	4,483
EXC	2005-01-03	2022-10-21	6,500	17.80	4,483
EXPD	2005-06-06	2022-10-21	6,346	17.37	4,377
EXPE	2005-08-09	2022-10-21	6,282	17.20	4,332
EXR	2005-01-03	2022-10-21	6,500	17.80	4,483
F	2005-01-03	2022-10-21	6,500	17.80	4,483
FAST	2005-01-03	2022-10-21	6,500	17.80	4,483
FCX	2005-01-03	2022-10-21	6,500	17.80	4,483
FDS	2007-05-01	2022-10-21	5,652	15.47	3,899
FDX	2005-01-03	2022-10-21	6,500	17.80	4,483
FE	2005-01-03	2022-10-21	6,500	17.80	4,483
FFIV	2005-01-03	2022-10-21	6,500	17.80	4,483
FHN	2005-01-03	2022-10-21	6,500	17.80	4,483
FIS	2006-02-01	2022-10-21	6,106	16.72	4,211
FISV	2005-01-03	2022-10-21	6,500	17.80	4,483
FL	2005-01-03	2022-10-21	6,500	17.80	4,483
FLEX	2005-01-03	2022-10-21	6,500	17.80	4,483
FLR	2005-01-03	2022-10-21	6,500	17.80	4,483
FLS	2005-01-03	2022-10-21	6,500	17.80	4,483
FMC	2005-01-03	2022-10-21	6,500	17.80	4,483
FOSL	2005-01-03	2022-10-21	6,500	17.80	4,483
FOX	2007-04-30	2022-10-21	5,653	15.48	3,878
FRT	2005-01-03	2022-10-21	6,500	17.80	4,483
FSLR	2006-11-17	2022-10-21	5,817	15.93	4,009
GD	2005-01-03	2022-10-21	6,500	17.80	4,483
GE	2005-01-03	2022-10-21	6,500	17.80	4,483
GHC	2007-05-16	2022-10-21	5,637	15.43	3,887
GILD	2005-01-03	2022-10-21	6,500	17.80	4,483
GIS	2005-01-03	2022-10-21	6,500	17.80	4,483
GL	2007-05-01	2022-10-21	5,652	15.47	3,897
GLW	2005-01-03	2022-10-21	6,500	17.80	4,483
GME	2005-10-10	2022-10-21	6,220	17.03	4,289
GNW	2005-01-03	2022-10-21	6,500	17.80	4,483
GOOG	2005-01-03	2022-10-21	6,500	17.80	4,483
GOOGL	2007-04-27	2022-10-21	5,656	15.49	3,878
GPC	2005-01-03	2022-10-21	6,500	17.80	4,483
GPN	2005-01-03	2022-10-21	6,500	17.80	4,483

GPS	2005-01-03	2022-10-21	6,500	17.80	4,483
GRMN	2005-01-03	2022-10-21	6,500	17.80	4,483
GS	2005-01-03	2022-10-21	6,500	17.80	4,483
GT	2005-01-03	2022-10-21	6,500	17.80	4,483
GWW	2005-01-03	2022-10-21	6,500	17.80	4,483
HAL	2005-01-03	2022-10-21	6,500	17.80	4,483
HAS	2005-01-03	2022-10-21	6,500	17.80	4,483
HBI	2006-09-06	2022-10-21	5,889	16.12	4,061
HD	2005-01-03	2022-10-21	6,500	17.80	4,483
HES	2007-04-27	2022-10-21	5,656	15.49	3,901
HIG	2005-01-03	2022-10-21	6,500	17.80	4,483
HOG	2006-09-01	2022-10-21	5,894	16.14	4,063
HOLX	2005-01-03	2022-10-21	6,500	17.80	4,483
HON	2005-01-03	2022-10-21	6,500	17.80	4,483
HP	2005-01-03	2022-10-21	6,500	17.80	4,483
HPQ	2005-01-03	2022-10-21	6,500	17.80	4,483
HRB	2005-01-03	2022-10-21	6,500	17.80	4,483
HRL	2005-01-03	2022-10-21	6,500	17.80	4,483
HSIC	2005-01-03	2022-10-21	6,500	17.80	4,483
HST	2006-04-18	2022-10-21	6,030	16.51	4,159
HSY	2005-01-03	2022-10-21	6,500	17.80	4,483
HUM	2005-01-03	2022-10-21	6,500	17.80	4,483
IBM	2005-01-03	2022-10-21	6,500	17.80	4,483
ICE	2005-11-16	2022-10-21	6,183	16.93	4,262
IDXX	2005-01-03	2022-10-21	6,500	17.80	4,483
IEX	2005-01-03	2022-10-21	6,500	17.80	4,483
IFF	2005-01-03	2022-10-21	6,500	17.80	4,483
ILMN	2005-01-03	2022-10-21	6,500	17.80	4,483
INCY	2005-01-03	2022-10-21	6,500	17.80	4,483
INTC	2005-01-03	2022-10-21	6,500	17.80	4,483
INTU	2005-01-03	2022-10-21	6,500	17.80	4,483
IP	2005-01-03	2022-10-21	6,500	17.80	4,483
IPG	2005-01-03	2022-10-21	6,500	17.80	4,483
IPGP	2006-12-13	2022-10-21	5,791	15.85	3,992
IRM	2005-01-03	2022-10-21	6,500	17.80	4,483
ISRG	2006-06-01	2022-10-21	5,986	16.39	4,128
IT	2005-01-03	2022-10-21	6,500	17.80	4,483
ITT	2007-05-01	2022-10-21	5,652	15.47	3,899
ITW	2007-04-30	2022-10-21	5,653	15.48	3,900
J	2007-05-02	2022-10-21	5,651	15.47	3,897
JBHT	2005-01-03	2022-10-21	6,500	17.80	4,483
JBL	2005-01-03	2022-10-21	6,500	17.80	4,483
JCI	2005-01-03	2022-10-21	6,500	17.80	4,483
JEF	2007-04-24	2022-10-21	5,659	15.49	3,904
JKHY	2005-01-03	2022-10-21	6,500	17.80	4,483
JNJ	2005-01-03	2022-10-21	6,500	17.80	4,483
JPM	2005-01-03	2022-10-21	6,500	17.80	4,483
JWN	2005-01-03	2022-10-21	6,500	17.80	4,483
K	2005-01-03	2022-10-21	6,500	17.80	4,483
KBH	2005-01-03	2022-10-21	6,500	17.80	4,483
KEY	2005-01-03	2022-10-21	6,500	17.80	4,483
KIM	2005-01-03	2022-10-21	6,500	17.80	4,483
KLAC	2005-01-03	2022-10-21	6,500	17.80	4,483
KMB	2005-01-03	2022-10-21	6,500	17.80	4,483
KMX	2005-01-03	2022-10-21	6,500	17.80	4,483
KO	2005-01-03	2022-10-21	6,500	17.80	4,483
KR	2005-01-03	2022-10-21	6,500	17.80	4,483
KSS	2005-01-03	2022-10-21	6,500	17.80	4,483
L	2007-04-27	2022-10-21	5,656	15.49	3,878
LBTYK	2005-09-08	2022-10-21	6,252	17.12	4,311
LDOS	2007-04-27	2022-10-21	5,656	15.49	3,900
LEG	2005-01-03	2022-10-21	6,500	17.80	4,483
LEN	2005-01-03	2022-10-21	6,500	17.80	4,483
LH	2005-01-03	2022-10-21	6,500	17.80	4,483
LHX	2007-05-01	2022-10-21	5,652	15.47	3,898
LKQ	2007-04-27	2022-10-21	5,656	15.49	3,900
LLY	2005-01-03	2022-10-21	6,500	17.80	4,483
LMT	2005-01-03	2022-10-21	6,500	17.80	4,483
LNC	2005-01-03	2022-10-21	6,500	17.80	4,483
LNT	2005-01-03	2022-10-21	6,500	17.80	4,483
LOW	2005-01-03	2022-10-21	6,500	17.80	4,483
LRCX	2005-01-03	2022-10-21	6,500	17.80	4,483
LUMN	2007-04-30	2022-10-21	5,653	15.48	3,899
LUV	2005-01-03	2022-10-21	6,500	17.80	4,483
LVS	2007-04-27	2022-10-21	5,656	15.49	3,901
LYV	2006-01-03	2022-10-21	6,135	16.80	4,231
M	2007-06-01	2022-10-21	5,621	15.39	3,877
MA	2006-05-25	2022-10-21	5,993	16.41	4,132
MAA	2005-01-03	2022-10-21	6,500	17.80	4,483
MAC	2005-01-03	2022-10-21	6,500	17.80	4,483
MAR	2005-01-03	2022-10-21	6,500	17.80	4,483

MAS	2005-01-03	2022-10-21	6,500	17.80	4,483
MAT	2005-01-03	2022-10-21	6,500	17.80	4,483
MBI	2005-01-03	2022-10-21	6,500	17.80	4,483
MCD	2005-01-03	2022-10-21	6,500	17.80	4,483
MCHP	2005-01-03	2022-10-21	6,500	17.80	4,483
MCK	2005-01-03	2022-10-21	6,500	17.80	4,483
MCO	2005-01-03	2022-10-21	6,500	17.80	4,483
MDLZ	2007-04-30	2022-10-21	5,653	15.48	3,899
MDT	2005-01-03	2022-10-21	6,500	17.80	4,483
MET	2005-01-03	2022-10-21	6,500	17.80	4,483
MGM	2005-06-01	2022-10-21	6,351	17.39	4,380
MHK	2005-01-03	2022-10-21	6,500	17.80	4,483
MKC	2005-01-03	2022-10-21	6,500	17.80	4,483
MKTX	2005-01-03	2022-10-21	6,500	17.80	4,483
MLM	2007-04-30	2022-10-21	5,653	15.48	3,900
MMC	2005-01-03	2022-10-21	6,500	17.80	4,483
MMM	2005-01-03	2022-10-21	6,500	17.80	4,483
MNST	2005-01-03	2022-10-21	6,500	17.80	4,483
MO	2005-01-03	2022-10-21	6,500	17.80	4,483
MOH	2007-01-03	2022-10-21	5,770	15.80	3,980
MPWR	2007-01-03	2022-10-21	5,770	15.80	3,980
MRK	2005-01-03	2022-10-21	6,500	17.80	4,483
MRO	2005-01-03	2022-10-21	6,500	17.80	4,483
MRVL	2005-01-03	2022-10-21	6,500	17.80	4,483
MS	2006-03-01	2022-10-21	6,078	16.64	4,192
MSCI	2007-11-15	2022-10-21	5,454	14.93	3,760
MSFT	2005-01-03	2022-10-21	6,500	17.80	4,483
MTB	2005-01-03	2022-10-21	6,500	17.80	4,483
MTD	2005-01-03	2022-10-21	6,500	17.80	4,483
MTW	2005-01-03	2022-10-21	6,500	17.80	4,483
MU	2005-01-03	2022-10-21	6,500	17.80	4,483
MUR	2005-01-03	2022-10-21	6,500	17.80	4,483
NBR	2005-11-03	2022-10-21	6,196	16.96	4,271
NDAQ	2005-02-10	2022-10-21	6,462	17.69	4,456
NDSN	2007-01-03	2022-10-21	5,770	15.80	3,980
NEE	2007-04-30	2022-10-21	5,653	15.48	3,899
NEM	2005-01-03	2022-10-21	6,500	17.80	4,483
NFLX	2005-01-03	2022-10-21	6,500	17.80	4,483
NI	2005-01-03	2022-10-21	6,500	17.80	4,483
NKE	2005-01-03	2022-10-21	6,500	17.80	4,483
NKTR	2005-01-03	2022-10-21	6,500	17.80	4,482
NLOK	2007-04-27	2022-10-21	5,656	15.49	3,900
NOC	2005-01-03	2022-10-21	6,500	17.80	4,483
NOV	2005-03-15	2022-10-21	6,429	17.60	4,434
NRG	2005-01-03	2022-10-21	6,500	17.80	4,483
NSC	2005-01-03	2022-10-21	6,500	17.80	4,483
NTAP	2005-01-03	2022-10-21	6,500	17.80	4,483
NUE	2005-01-03	2022-10-21	6,500	17.80	4,483
NVDA	2005-01-03	2022-10-21	6,500	17.80	4,483
NWL	2005-01-03	2022-10-21	6,500	17.80	4,483
NYT	2005-01-03	2022-10-21	6,500	17.80	4,483
ODFL	2005-01-03	2022-10-21	6,500	17.80	4,483
ODP	2005-01-03	2022-10-21	6,500	17.80	4,483
OI	2005-01-03	2022-10-21	6,500	17.80	4,483
OKE	2005-01-03	2022-10-21	6,500	17.80	4,483
OMC	2005-01-03	2022-10-21	6,500	17.80	4,483
ORCL	2005-01-03	2022-10-21	6,500	17.80	4,483
ORLY	2005-01-03	2022-10-21	6,500	17.80	4,483
OXY	2005-01-03	2022-10-21	6,500	17.80	4,483
PAR	2007-05-16	2022-10-21	5,637	15.43	3,884
PARA	2007-04-30	2022-10-21	5,653	15.48	3,899
PAYX	2005-06-06	2022-10-21	6,346	17.37	4,377
PBI	2005-01-03	2022-10-21	6,500	17.80	4,483
PCAR	2005-01-03	2022-10-21	6,500	17.80	4,483
PCG	2005-01-03	2022-10-21	6,500	17.80	4,483
PDCO	2005-01-03	2022-10-21	6,500	17.80	4,483
PEAK	2007-04-30	2022-10-21	5,653	15.48	3,880
PEG	2005-01-03	2022-10-21	6,500	17.80	4,483
PENN	2007-01-03	2022-10-21	5,770	15.80	3,980
PEP	2005-01-03	2022-10-21	6,500	17.80	4,483
PFE	2005-01-03	2022-10-21	6,500	17.80	4,483
PFG	2005-01-03	2022-10-21	6,500	17.80	4,483
PG	2005-01-03	2022-10-21	6,500	17.80	4,483
PGR	2007-04-26	2022-10-21	5,657	15.49	3,902
PH	2005-01-03	2022-10-21	6,500	17.80	4,483
PHM	2005-01-03	2022-10-21	6,500	17.80	4,483
PKG	2005-01-03	2022-10-21	6,500	17.80	4,483
PKI	2005-01-03	2022-10-21	6,500	17.80	4,483
PLD	2007-04-27	2022-10-21	5,656	15.49	3,901
PNC	2005-01-03	2022-10-21	6,500	17.80	4,483
PNW	2005-01-03	2022-10-21	6,500	17.80	4,483

POOL	2005-01-03	2022-10-21	6,500	17.80	4,483
PPG	2005-01-03	2022-10-21	6,500	17.80	4,483
PPL	2005-01-03	2022-10-21	6,500	17.80	4,483
PRGO	2005-01-03	2022-10-21	6,500	17.80	4,483
PRU	2005-01-03	2022-10-21	6,500	17.80	4,483
PSA	2005-01-03	2022-10-21	6,500	17.80	4,483
PTC	2007-04-27	2022-10-21	5,656	15.49	3,901
PVH	2005-01-03	2022-10-21	6,500	17.80	4,483
PWR	2005-01-03	2022-10-21	6,500	17.80	4,483
PXD	2005-01-03	2022-10-21	6,500	17.80	4,483
QCOM	2005-01-03	2022-10-21	6,500	17.80	4,483
R	2005-01-03	2022-10-21	6,500	17.80	4,483
RCL	2005-01-03	2022-10-21	6,500	17.80	4,483
RE	2005-01-03	2022-10-21	6,500	17.80	4,483
REG	2005-01-03	2022-10-21	6,500	17.80	4,483
REGN	2005-01-03	2022-10-21	6,500	17.80	4,483
RF	2005-01-03	2022-10-21	6,500	17.80	4,483
RHI	2005-01-03	2022-10-21	6,500	17.80	4,483
RJF	2005-01-03	2022-10-21	6,500	17.80	4,483
RL	2005-01-03	2022-10-21	6,500	17.80	4,483
RMD	2005-01-03	2022-10-21	6,500	17.80	4,483
ROK	2005-01-03	2022-10-21	6,500	17.80	4,483
ROL	2005-01-03	2022-10-21	6,500	17.80	4,483
ROP	2007-04-30	2022-10-21	5,653	15.48	3,900
ROST	2005-01-03	2022-10-21	6,500	17.80	4,483
RRC	2005-01-03	2022-10-21	6,500	17.80	4,483
RSG	2005-01-03	2022-10-21	6,500	17.80	4,483
RTX	2007-04-27	2022-10-21	5,656	15.49	3,900
SANM	2007-04-30	2022-10-21	5,653	15.48	3,881
SBAC	2005-01-03	2022-10-21	6,500	17.80	4,483
SBUX	2005-01-03	2022-10-21	6,500	17.80	4,483
SEE	2005-01-03	2022-10-21	6,500	17.80	4,483
SHW	2005-01-03	2022-10-21	6,500	17.80	4,483
SIG	2005-01-03	2022-10-21	6,500	17.80	4,483
SIRI	2005-01-03	2022-10-21	6,500	17.80	4,483
SITC	2007-05-01	2022-10-21	5,652	15.47	3,898
SIVB	2006-03-01	2022-10-21	6,078	16.64	4,192
SJM	2005-01-03	2022-10-21	6,500	17.80	4,483
SLB	2005-01-03	2022-10-21	6,500	17.80	4,483
SLG	2005-01-03	2022-10-21	6,500	17.80	4,483
SNA	2005-01-03	2022-10-21	6,500	17.80	4,483
SNPS	2005-01-03	2022-10-21	6,500	17.80	4,483
SO	2005-01-03	2022-10-21	6,500	17.80	4,483
SPG	2005-01-03	2022-10-21	6,500	17.80	4,483
SPGI	2007-04-27	2022-10-21	5,656	15.49	3,900
SRCL	2005-01-03	2022-10-21	6,500	17.80	4,483
SRE	2005-01-03	2022-10-21	6,500	17.80	4,483
SSP	2005-01-03	2022-10-21	6,500	17.80	4,483
STT	2005-01-03	2022-10-21	6,500	17.80	4,483
STX	2005-01-03	2022-10-21	6,500	17.80	4,483
STZ	2005-01-03	2022-10-21	6,500	17.80	4,483
SWK	2005-01-03	2022-10-21	6,500	17.80	4,483
SWKS	2005-01-03	2022-10-21	6,500	17.80	4,483
SWN	2005-01-03	2022-10-21	6,500	17.80	4,483
SYK	2005-01-03	2022-10-21	6,500	17.80	4,483
SYY	2005-01-03	2022-10-21	6,500	17.80	4,483
T	2006-01-03	2022-10-21	6,135	16.80	4,231
TAP	2005-03-01	2022-10-21	6,443	17.64	4,444
TDC	2007-10-01	2022-10-21	5,499	15.06	3,793
TDG	2006-03-15	2022-10-21	6,064	16.60	4,182
TDY	2007-01-03	2022-10-21	5,770	15.80	3,980
TECH	2007-01-03	2022-10-21	5,770	15.80	3,980
TER	2005-01-03	2022-10-21	6,500	17.80	4,483
TEX	2005-01-03	2022-10-21	6,500	17.80	4,483
TFC	2007-04-30	2022-10-21	5,653	15.48	3,877
TFX	2005-01-03	2022-10-21	6,500	17.80	4,483
TGNA	2007-04-30	2022-10-21	5,653	15.48	3,899
TGT	2005-01-03	2022-10-21	6,500	17.80	4,483
THC	2005-01-03	2022-10-21	6,500	17.80	4,483
TJX	2005-01-03	2022-10-21	6,500	17.80	4,483
TMO	2005-01-03	2022-10-21	6,500	17.80	4,483
TMUS	2007-05-16	2022-10-21	5,637	15.43	3,887
TPR	2007-04-27	2022-10-21	5,656	15.49	3,900
TRMB	2007-01-03	2022-10-21	5,770	15.80	3,980
TROW	2005-01-03	2022-10-21	6,500	17.80	4,483
TRV	2007-02-27	2022-10-21	5,715	15.65	3,943
TSCO	2005-01-03	2022-10-21	6,500	17.80	4,483
TSN	2005-01-03	2022-10-21	6,500	17.80	4,483
TT	2007-04-30	2022-10-21	5,653	15.48	3,882
TTWO	2005-01-03	2022-10-21	6,500	17.80	4,483
TUP	2005-01-03	2022-10-21	6,500	17.80	4,483

TXN	2005-01-03	2022-10-21	6,500	17.80	4,483
TXT	2005-01-03	2022-10-21	6,500	17.80	4,483
TYL	2007-01-03	2022-10-21	5,770	15.80	3,980
UAA	2007-05-14	2022-10-21	5,639	15.44	3,889
UDR	2005-01-03	2022-10-21	6,500	17.80	4,483
UHS	2005-01-03	2022-10-21	6,500	17.80	4,483
ULTA	2007-10-25	2022-10-21	5,475	14.99	3,775
UNH	2005-01-03	2022-10-21	6,500	17.80	4,483
UNM	2005-01-03	2022-10-21	6,500	17.80	4,483
UNP	2005-01-03	2022-10-21	6,500	17.80	4,483
UPS	2005-01-03	2022-10-21	6,500	17.80	4,483
URBN	2005-01-03	2022-10-21	6,500	17.80	4,483
URI	2005-01-03	2022-10-21	6,500	17.80	4,483
USB	2005-01-03	2022-10-21	6,500	17.80	4,483
VFC	2005-01-03	2022-10-21	6,500	17.80	4,483
VIAV	2007-04-27	2022-10-21	5,656	15.49	3,879
VLO	2005-01-03	2022-10-21	6,500	17.80	4,483
VMC	2007-11-19	2022-10-21	5,450	14.92	3,758
VNO	2005-01-03	2022-10-21	6,500	17.80	4,483
VRSN	2005-01-03	2022-10-21	6,500	17.80	4,483
VRTX	2005-01-03	2022-10-21	6,500	17.80	4,483
VTR	2005-01-03	2022-10-21	6,500	17.80	4,483
VZ	2005-01-03	2022-10-21	6,500	17.80	4,483
WAB	2005-01-03	2022-10-21	6,500	17.80	4,483
WAT	2005-01-03	2022-10-21	6,500	17.80	4,483
WDC	2005-01-03	2022-10-21	6,500	17.80	4,483
WEC	2005-01-03	2022-10-21	6,500	17.80	4,483
WELL	2007-04-30	2022-10-21	5,653	15.48	3,880
WFC	2005-01-03	2022-10-21	6,500	17.80	4,483
WHR	2005-01-03	2022-10-21	6,500	17.80	4,483
WM	2007-04-30	2022-10-21	5,653	15.48	3,900
WMB	2005-01-03	2022-10-21	6,500	17.80	4,483
WMT	2005-01-03	2022-10-21	6,500	17.80	4,483
WOR	2005-01-03	2022-10-21	6,500	17.80	4,483
WRB	2007-04-27	2022-10-21	5,656	15.49	3,900
WST	2007-01-03	2022-10-21	5,770	15.80	3,980
WU	2006-10-02	2022-10-21	5,863	16.05	4,043
WY	2005-01-03	2022-10-21	6,500	17.80	4,483
WYNN	2005-01-03	2022-10-21	6,500	17.80	4,483
X	2005-01-03	2022-10-21	6,500	17.80	4,483
XEL	2005-01-03	2022-10-21	6,500	17.80	4,483
XOM	2005-01-03	2022-10-21	6,500	17.80	4,483
XRAY	2005-01-03	2022-10-21	6,500	17.80	4,483
XRX	2005-01-03	2022-10-21	6,500	17.80	4,483
YUM	2005-01-03	2022-10-21	6,500	17.80	4,483
ZBH	2007-04-30	2022-10-21	5,653	15.48	3,899
ZBRA	2005-01-03	2022-10-21	6,500	17.80	4,483
ZION	2005-01-03	2022-10-21	6,500	17.80	4,483

Table 30: Data description of all 478 stocks included in the final panel data set. Illustrated are the data sample start and end dates, the number of days and years, and the number of trading days, respectively.

References

Aït-Sahalia, Y., Mykland, P. A., & Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *The Review of Financial Studies*, *18*(2), 351–416. https://www.jstor.org/stable/3598041

Akgiray, V. (1989). Conditional heteroscedasticity in time series of stock returns: Evidence and forecasts. *The Journal of Business*, *62*(1), 55–80. https://www.jstor.org/stable/2353123

Alizadeh, S., Brandt, M. W., & Diebold, F. X. (2002). Range-based estimation of stochastic volatility models. *The Journal of Finance*, *57*(3), 1047–1091. https://doi.org/10.1111/1540-6261.00454

Alsubaie, A., & Najand, M. (2009). Trading volume, time-varying conditional volatility, and asymmetric volatility spillover in the saudi stock market. *Journal of Multinational Financial Management*, *19*(2), 139–159. https://doi.org/10.1016/j.mulfin.2008.09.002

Andersen, T. G., Bollerslev, T., & Meddahi, N. (2011). Realized volatility forecasting and market microstructure noise. *Journal of Econometrics*, *160*(1), 220–234. https://doi.org/10.1016/j.jeconom.2010.03.032

Audrino, F., & Knaus, S. D. (2016). Lassoing the HAR model: A model selection perspective on realized volatility dynamics. *Econometric Reviews*, *35*(8), 1485–1521. https://doi.org/10.1080/07474938.2015.1092801

Audrino, F., Sigrist, F., & Ballinari, D. (2020). The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting*, *36*(2), 334–357. https://doi.org/10.1016/j.ijforecast.2019.05.010

Baillie, R. T., & Bollerslev, T. (1992). Prediction in dynamic models with time-dependent conditional variances. *Journal of Econometrics*, *52*(1), 91–113. https://doi.org/10.1016/0304-4076(92)90066-Z

Bank for International Settlements, B. (2022). OTC derivatives statistics at end-june 2022. Retrieved March 21, 2023, from https://www.bis.org/publ/otc_hy2211.html

Barndorff-Nielsen, O. E., & Shephard, N. (2002). Estimating quadratic variation using realized variance. *Journal of Applied Econometrics*, *17*(5), 457–477. https://www.jstor.org/stable/4129267

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, *81*(3), 637–654. https://www.jstor.org/stable/1831029

Black, F. (1976). Studies of stock market volatility changes. *Proceedings of the American Statistical Association, Business & Economic Statistics Section, 1976*. https://cir.nii.ac.jp/crid/1570009749981528192

Bollerslev, T. (2006). Leverage and volatility feedback effects in high-frequency data. *Journal of Financial Econometrics*, *4*(3), 353–384. https://doi.org/10.1093/jjfinec/nbj014

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Bollerslev, T., Hood, B., Huss, J., & Pedersen, L. H. (2018). Risk everywhere: Modeling and managing volatility. *The Review of Financial Studies*, *31*(7), 2729–2773. https://doi.org/10.1093/rfs/hhy041

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140. https://doi.org/10.1023/A:1018054314350

Bucci, A. (2020). Realized volatility forecasting with neural networks. *Journal of Financial Econometrics*, *18*(3), 502–531. https://doi.org/10.1093/jjfinec/nbaa008

Campbell, J., & Hentschel, L. (1991). *No news is good news: An asymmetric model of changing volatility in stock returns*. National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w3742

Caporin, M., & Poli, F. (2017). Building news measures from textual data and an application to volatility forecasting. *Econometrics*, *5*(3), 35. https://doi.org/10.3390/econometrics5030035

Castanias, R. P. (1979). Macroinformation and the variability of stock market prices. *The Journal of Finance*, *34*(2), 439–450. https://doi.org/10.2307/2326984

Chen, C.-H., Yu, W.-C., & Zivot, E. (2012). Predicting stock volatility using after-hours information: Evidence from the NASDAQ actively traded stocks. *International Journal of Forecasting*, *28*(2), 366–383. https://doi.org/10.1016/j.ijforecast.2011.04.005

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Christensen, K., Siggaard, M., & Veliyev, B. (2021). A machine learning approach to volatility forecasting. Retrieved January 10, 2023, from https://papers.ssrn.com/abstract=3766999

Christie, A. (1982). The stochastic behavior of common stock variances value, leverage and interest rate effects. *Journal of Financial Economics*, *10*(4), 407–432. https://doi.org/10.1016/0304-405X(82)90018-6

Chuang, C.-C., Kuan, C.-M., & Lin, H.-Y. (2009). Causality in quantiles and dynamic stock return–volume relations. *Journal of Banking & Finance*, *33*(7), 1351–1360. https://doi.org/10.1016/j.jbankfin.2009.02.013

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, *1*(2), 223–236. https://doi.org/10.1080/713665670

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, *7*(2), 174–196. https://doi.org/10.1093/jjfinec/nbp001

Corsi, F., Mittnik, S., Pigorsch, C., & Pigorsch, U. (2008). The volatility of realized volatility. *Econometric Reviews*, *27*(1), 46–78. https://doi.org/10.1080/07474930701853616

Corsi, F., & Renò, R. (2012). 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. Retrieved March 10, 2023, from https://www.jstor.org/stable/23243735

Cox, J. C., Ingersoll, J. E., & Ross, S. A. (1985). An intertemporal general equilibrium model of asset prices. *Econometrica*, *53*(2), 363–384. https://doi.org/10.2307/1911241

Deng, H., Zhou, Y., Wang, L., & Zhang, C. (2021). Ensemble learning for the early prediction of neonatal jaundice with genetic features. *BMC Medical Informatics and Decision Making*, *21*(1), 338. https://doi.org/10.1186/s12911-021-01701-9

Derman, E., Kani, I., & Zou, J. Z. (1996). The local volatility surface: Unlocking the information in index option prices. *Financial Analysts Journal*, *52*(4), 25–36. Retrieved January 9, 2023, from https://www.jstor.org/stable/4479931

Ding, S., Cui, T., & Zhang, Y. (2022). Futures volatility forecasting based on big data analytics with incorporating an order imbalance effect. *International Review of Financial Analysis*, *83*, 102255. https://doi.org/10.1016/j.irfa.2022.102255

Donaldson, R. G., & Kamstra, M. (1997). An artificial neural network-GARCH model for international stock return volatility. *Journal of Empirical Finance*, *4*(1), 17–46. https://doi.org/10.1016/S0927-5398(96)00011-4

Dupire, B. (1994). Pricing with a smile. *Risk*, *7*(1), 18–20.

EarningsDates. (2023). *Historic stock earnings dates - CSV earnings data parsed directly from the SEC / EDGAR*. Retrieved January 12, 2023, from https://earnings-dates.com/

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, *50*(4), 987–1007. https://doi.org/10.2307/1912773

Engle, R. F., & Rosenberg, J. V. (1995). GARCH gamma. https://doi.org/10.3386/w5128

Fama, E. F. (1965). The behavior of stock-market prices. *The Journal of Business*, *38*(1), 34–105. Retrieved January 10, 2023, from https://www.jstor.org/stable/2350752

Fernandes, M., Medeiros, M. C., & Scharth, M. (2014). Modeling and predicting the CBOE market volatility index. *Journal of Banking & Finance*, *40*, 1–10. https://doi.org/10.1016/j.jbankfin.2013.11.004

FirstRate Data. (2023). *S&P 500 historical intraday prices bundle*. Retrieved January 5, 2023, from https://firstratedata.com/b/1/sp500-historical-intraday-stocks-bundle

Forsberg, L., & Ghysels, E. (2007). Why do absolute returns predict volatility so well? *Journal of Financial Econometrics*, *5*(1), 31–67. https://doi.org/10.1093/jjfinec/nbl010

Foschi, P., & Pascucci, A. (2008). Path dependent volatility. *Decisions in Economics and Finance*, *31*(1), 13–32. https://doi.org/10.1007/s10203-007-0076-6

French, K. R., Schwert, G., & Stambaugh, R. F. (1987). Expected stock returns and volatility. *Journal of Financial Economics*, *19*(1), 3–29. https://doi.org/10.1016/0304-405X(87)90026-2

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. Retrieved April 4, 2023, from https://www.jstor.org/stable/2699986

Gençay, R., Gradojevic, N., Selçuk, F., & Whitcher, B. (2010). Asymmetry of information flow between volatilities across time scales. *Quantitative Finance*, *10*(8), 895–915. https://doi.org/10.1080/14697680903460143

Ghysels, E., & Sinko, A. (2011). Volatility forecasting and microstructure noise. *Journal of Econometrics*, *160*(1), 257–271. https://doi.org/10.1016/j.jeconom.2010.03.035

Giordani, P. (2021). Smartboost learning for tabular data. https://doi.org/10.2139/ssrn.3975543

Greenwell, B. M., Boehmke, B. C., & McCarthy, A. J. (2018). A simple and effective model-based variable importance measure. Retrieved April 5, 2023, from http://arxiv.org/abs/1805.04755

Hagan, P. S., Kumar, D., Andrew S., L., & Woodward, D. E. (2002). Managing smile risk. Retrieved January 10, 2023, from https://archive.ph/jrQyW

Hansen, P. R., & Lunde, A. (2005). A realized variance for the whole day based on intermittent high-frequency data. *Journal of Financial Econometrics*, *3*(4), 525–554. https://doi.org/10.1093/jjfinec/nbi028

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* Springer Science & Business Media.

Heston, S. L. (1993). Closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, *6*(2), 327–343. Retrieved January 9, 2023, from https://academic.oup.com/rfs/article/6/2/327/1574747?login=true

Hibbert, A. M., Daigler, R. T., & Dupoyet, B. (2008). A behavioral explanation for the negative asymmetric return–volatility relation. *Journal of Banking & Finance*, *32*(10), 2254–2266. https://doi.org/10.1016/j.jbankfin.2007.12.046

Hillebrand, E., & Medeiros, M. (2010). The benefits of bagging for forecast models of realized volatility. *Econometric Reviews*, *29*(5), 571–593. https://doi.org/10.1080/07474938.2010.481554

Hull, J. (2015). *Options, futures, and other derivatives* (Ninth edition). Pearson.

Hull, J., & White, A. (1987). The pricing of options on assets with stochastic volatilities. *The Journal of Finance*, *42*(2), 281–300. https://doi.org/10.2307/2328253

Kabir, M. H., & Hassan, M. K. (2005). The near-collapse of LTCM, US financial stock returns, and the fed. *Journal of Banking & Finance*, *29*(2), 441–460. https://doi.org/10.1016/j.jbankfin.2004.05.014

Karpoff, J. M. (1987). The relation between price changes and trading volume: A survey. *The Journal of Financial and Quantitative Analysis*, *22*(1), 109–126. https://doi.org/10.2307/2330874

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision

tree. *Advances in Neural Information Processing Systems*, *30*. Retrieved April 5, 2023, from https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Li, S. Z., & Tang, Y. (2022). Automated risk forecasting. https://doi.org/10.2139/ssrn.3776915

LightGBM. (2023a). *Leaf-wise. Reprinted from LightGBM documentation*. Retrieved April 5, 2023, from https://lightgbm.readthedocs.io/en/latest/_images/leaf-wise.png

LightGBM. (2023b). *Level-wise. Reprinted from LightGBM documentation*. Retrieved April 5, 2023, from https://lightgbm.readthedocs.io/en/latest/_images/level-wise.png

Liu, H., & Loewenstein, M. (2009). Market crashes, correlated illiquidity, and portfolio choice. Retrieved January 10, 2023, from https://www.semanticscholar.org/paper/Market-Crashes%2C-Correlated-Illiquidity%2C-and-%5CFlight-Liu-Loewenstein/3817d126357bb05855cc486d34c8ce3ee18cac27

Liu, L. Y., Patton, A. J., & Sheppard, K. (2015). Does anything beat 5-minute RV? a comparison of realized measures across multiple asset classes. *Journal of Econometrics*, *187*(1), 293–311. https://doi.org/10.1016/j.jeconom.2015.02.008

Liu, Z. (2022). Stock volatility prediction using LightGBM based algorithm. *2022 International Conference on Big Data, Information and Computer Network (BDICN)*, 283–286. https://doi.org/10.1109/BDICN55575.2022.00061

Louhichi, W. (2011). What drives the volume–volatility relationship on euronext paris? *International Review of Financial Analysis*, *20*(4), 200–206. https://doi.org/10.1016/j.irfa.2011.03.001

Lowenstein, R. (2001). *When genius failed: The rise and fall of long-term capital management* (Reprint edition). Random House Trade Paperbacks.

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Luong, C., & Dokuchaev, N. (2018). Forecasting of realised volatility with the random forests algorithm. *Journal of Risk and Financial Management*, *11*(4), 61. https://doi.org/10.3390/jrfm11040061

Mandelbrot, B. (1963). New methods in statistical economics. *Journal of Political Economy*, *71*(5), 421–440. Retrieved January 10, 2023, from https://www.jstor.org/stable/1829014

Masset, P. (2011). Volatility stylized facts. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1804070

Mei, D., Liu, J., Ma, F., & Chen, W. (2017). Forecasting stock market volatility: Do realized skewness and kurtosis help? *Physica A: Statistical Mechanics*

and its Applications, *481*, 153–159. https://doi.org/10.1016/j.physa.2017.04.020

Merton, R. C. (1973). Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, *4*(1), 141–183. https://doi.org/10.2307/3003143

Mittnik, S., Robinzonov, N., & Spindler, M. (2015). Stock market volatility: Identifying major drivers and the nature of their impact. *Journal of Banking & Finance*, *58*, 1–14. https://doi.org/10.1016/j.jbankfin.2015.04.003

Orlando, G., & Taglialatela, G. (2017). A review on implied volatility calculation. *Journal of Computational and Applied Mathematics*, *320*, 202–220. https://doi.org/10.1016/j.cam.2017.02.002

Pindyck, R. (1983). *Risk, inflation, and the stock market*. National Bureau of Economic Research. https://doi.org/10.3386/w1186

Poterba, J., & Summers, L. (1984). *The persistence of volatility and stock market fluctuations*. National Bureau of Economic Research. https://doi.org/10.3386/w1462

Rachev, S. T. (2003). *Handbook of heavy tailed distributions in finance*. Elsevier.

Rahimikia, E., & Poon, S.-H. (2020). Machine learning for realised volatility forecasting. https://doi.org/10.2139/ssrn.3707796

Ren, Y., Madan, D., & Qian, M. Q. (2007). Calibrating and pricing with embedded local volatility models. *Risk.net*. Retrieved January 9, 2023, from https://www.risk.net/node/1500232

RiskMetrics. (1996). RiskMetrics technical document.

Shephard, N. (2005). *Stochastic volatility: Selected readings*. Oxford University Press.

Stoll, H. R., & Whaley, R. E. (1990). The dynamics of stock index and stock index futures returns. *Journal of Financial and Quantitative analysis*, *25*(4), 441–468.

Teller, A., Pigorsch, U., & Pigorsch, C. (2022). Short- to long-term realized volatility forecasting using extreme gradient boosting. https://doi.org/10.2139/ssrn.4267541

Teyssière, G., & Kirman, A. P. (Eds.). (2007). *Long memory in economics*. Springer.

Turner, A. L., & Weigel, E. J. (1992). Daily stock market volatility: 1928-1989. *Management Science*, *38*(11), 1586–1609. Retrieved January 10, 2023, from https://www.jstor.org/stable/2632471

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics*, *5*(2), 177–188. https://doi.org/10.1016/0304-405X(77)90016-2

VIX index. (2023). Retrieved January 6, 2023, from https://www.cboe.com/tradable_products/vix/

Wing-Yi Chio, S., Li, Y., & JingRan Yang, R. (2022). Realized volatility prediction. *2021 2nd European Symposium on Software Engineering*, 129–135. https://doi.org/10.1145/3501774.3501793

Yahoo Finance. (2023). CBOE volatility index (ˆVIX) historical data - Yahoo Finance. Retrieved January 5, 2023, from https://consent.yahoo.com/v2/collectConsent?sessionId=3_cc-session_36285256-323a-4e88-bd52-0cd9c588dc06

Ypma, T. J. (1995). Historical development of the newton–raphson method. *SIAM Review*, *37*(4), 531–551. https://doi.org/10.1137/1037125

Zhu, X., Zhang, H., & Zhong, M. (2017). Volatility forecasting using high frequency data: The role of after-hours information and leverage effects. *Resources Policy*, *54*, 58–70. https://doi.org/10.1016/j.resourpol.2017.09.006