Quantitative Finance

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rquf20

Predicting VIX with adaptive machine learning

Yunfei Bai & Charlie X. Cai

To cite this article: Yunfei Bai & Charlie X. Cai (2024) Predicting VIX with adaptive machine learning, Quantitative Finance, 24:12, 1857-1873, DOI: 10.1080/14697688.2024.2439458

To link to this article: https://doi.org/10.1080/14697688.2024.2439458

Quantitative Finance, 2024
Vol. 24, No. 12, 1857–1873, https://doi.org/10.1080/14697688.2024.2439458

Predicting VIX with adaptive machine learning

YUNFEI BAI †† and CHARLIE X. CAI ‡*

†AI/ML and Big Data Consultant, Seattle, USA
‡Finance, Liverpool University School of Management, University of Liverpool, Liverpool, UK

(*Received 19 January 2024; accepted 28 November 2024*)

This paper investigates the predictability of the CBOE Volatility Index (VIX) and explores the sources of its predictability using machine learning (ML) techniques. We establish that daily VIX can be predicted with higher accuracy than previously documented, yielding forecasts of significant economic value. Our analysis underscores the efficacy of dynamic training, nonlinear methods and a comprehensive set of economic variables in predicting VIX trends. We identify the weekly jobless claim data as a pivotal variable, revealing its substantial influence on market volatility, an area not extensively explored in prior research. While accurately forecasting VIX spikes poses a challenge, our algorithms demonstrate remarkable adaptability to new data, thereby significantly enhancing the resilience of trading strategies. This research not only contributes to the understanding of VIX predictability but also offers valuable insights for the development of more robust quantitative investment and risk management strategies.

Keywords: Machine learning; AutoML; Explainable AI; VIX; Predictability; Forecasting; Quantitative trading; Big data; S&P 500; Futures; US markets

JEL Classification: G0, G17, C52, C55, C58

1. Introduction

The CBOE Volatility Index (VIX), often referred to as the 'fear index' (Whaley 2000), is a key forward-looking indicator for market participants and policymakers. Despite its prominence, VIX predictability remains challenging due to traditional reliance on limited predictors and linear models. This paper leverages recent Machine Learning (ML) advancements to address these limitations by exploring the predictability of the VIX using a wide range of economic indicators, focusing on forecasting accuracy and economic relevance, and examining the sources and constraints of this predictability.

While finance-ML literature has largely focused on cross-sectional return forecasting, research on time series volatility forecasting remains limited. Our study bridges this gap by predicting the VIX's directional movement for the following day, aligning with binary investment decisions (long or short). We compile 278 features across 14 categories from Bloomberg, including global markets, macroeconomic data, and seasonality factors, ensuring real-time data availability without look-back bias.

We employ various ML algorithms, ranging from Naïve Bayes and Logistic Regression to Decision Trees, Random Forests, Adaptive Boosting (AB), and Multi-Layer Perceptron, alongside an ensemble model. A novel cross-validation strategy maintains the time-series integrity of VIX data (Bergmeir *et al.* 2018). Our study segments the data into three periods: In-Sample (pre-2009) for training, Out-of-Sample (late 2009) for validation, and Implementation (2010–2020) for testing.

Key findings reveal that Adaptive Boosting (AB) achieved the highest validation accuracy (68.2%), outperforming other models and demonstrating resilience against overfitting. Moreover, AB's forecasts delivered an annualized return of 225% in a simulated long/short VIX investment strategy, with a Sharpe ratio of 1.7, underscoring its practical value. Our economic evaluation shows that VIX forecasts have significant applications in valuation and risk management models. Additionally, the models effectively adjusted to unexpected market events, such as the 2010 Flash Crash and 2016 Brexit referendum, recovering losses more efficiently than non-model strategies.

Variable importance analysis highlights the US weekly jobless report as the most influential predictor, followed by seasonality factors such as days until VIX futures expiration. Our findings suggest that relying solely on VIX's historical data overlooks critical economic information. The inclusion of a broader set of features enables a deeper understanding of

*Corresponding author. Email: busxc@liverpool.ac.uk

VIX predictability, particularly in capturing market behaviors like momentum and reversal patterns.

Further tests show that dynamic retraining and balanced sampling significantly improve model performance, particularly for predicting larger market movements. Expanding to a four-category model (up-small, up-big, down-small, down-big) maintains accuracy. Our ML framework's adaptability and automation offer valuable insights for future financial forecasting applications.

Our research contributes significantly in two areas. First, in finance literature, we demonstrate the superiority of machine learning (ML) over traditional methods like Logistic Regression and HAR for volatility forecasting. This extends the work of Konstantinidi *et al.* (2008), Paye (2012), and Fernandes *et al.* (2014), showing that ML's ability to include a broader range of economic variables provides deeper insights into the relationship between macroeconomic factors and volatility. The extended historical dataset also enhances cross-validation and out-of-sample testing.

Second, our study offers practical value for quantitative investment strategies and risk management by accurately predicting VIX direction. This helps investors make informed entry and exit decisions and allows for better risk management through volatility forecasts. The model is also highly useful for derivatives trading, as it anticipates market movements and quickly adapts to new information, making it valuable for both short-term and long-term investors.

Additionally, our research offers insights for future ML applications in financial forecasting. We propose an adaptive learning framework featuring Automated Machine Learning (AutoML) and Hyperparameter Optimization (HPO) for continuous, out-of-sample implementation. This framework reduces manual intervention and ensures the model evolves with changing market conditions, serving as a reference for future ML studies in finance.

The structure of this paper is outlined as follows: Section 2 provides an overview of existing literature on volatility forecasting, setting the context for our study. In Section 3, we detail our research design, outlining the methodologies and approaches employed. The empirical findings are presented in Sections 4, 5, and 6, where we delve into the forecasting performance, conduct an economic evaluation, and explore the sources of predictability, respectively. Finally, Section 7 offers concluding remarks, summarizing the key insights and implications of our research.

2. Related literature

2.1. *Machine learning applications in economics and finance*

The rapid advancements in machine learning (ML) and computing power have revolutionized fields like economics and finance, enabling sophisticated models to process vast data and uncover patterns missed by traditional methods. This review engages with recent developments in ML applications in these areas. Breiman's (2001) foundational work on ensemble learning, particularly the random forest algorithm, has been pivotal, making ensemble methods essential due to their robustness and accuracy in fields like financial analytics.

In finance, BlackRock's 2019 report highlights ML's transformative impact, showing how algorithms enhance decision-making, optimize portfolios, and manage risks by analyzing large datasets and predicting market trends more accurately than traditional methods (BackRock 2019). The CFA Institute (2020) also emphasizes ML's role in asset management, covering its use in automated trading, fraud detection, credit scoring, and personalized advice, while addressing ethical and regulatory concerns.

The integration of ML with economics and finance has opened new research avenues, enabling the analysis of economic indicators and financial trends through models capable of handling complex data. This paper explores how advanced ML techniques, combined with a wide set of economic variables, can improve our understanding of the VIX's predictability, a critical financial indicator.

2.2. *Historical and realized volatility forecasting*

The foundational models in volatility forecasting are the GARCH family models, introduced by Engle (1982) and Bollerslev (1986). Andersen *et al.* (2003) later showed that multivariate realized volatility modeling outperforms the GARCH and stochastic volatility models in out-of-sample forecasts. Corsi's (2009) heterogeneous autoregressive (HAR) model, which captures volatility persistence over daily, weekly, and monthly horizons, has become a popular benchmark due to its simplicity and effectiveness in replicating empirical volatility patterns.

Many studies have integrated GARCH or HAR with nonlinear methods to improve forecasting accuracy. Kristjanpoller and Minutolo (2018), Maciel *et al.* (2016), and Psaradellis and Sermpinis (2016) contributed to this area. Donaldson and Kamstra (1997) found that an Artificial Neural Network-GARCH model outperformed traditional models like GARCH and EGARCH in forecasting stock return volatility. More recently, Bucci (2020) demonstrated that LSTM Recursive Neural Networks (RNNs) can exceed linear models in out-of-sample forecasts of monthly realized volatility. However, most studies have focused primarily on time series methods, with less attention to broader economic factors.

2.3. *Implied volatility and VIX forecasting*

Implied volatility has been a key measure of market expectations since the VIX index was introduced by CBOE in 1993. Known as the 'fear index,' the VIX is a benchmark for U.S. equity market volatility, reflecting the 30-day expected volatility based on SPX options prices. Early work by Hamid and Iqbal (2004) showed neural networks outperform implied volatility in predicting S&P 500 futures prices, though our focus is on forecasting the VIX itself. Konstantinidi *et al.* (2008) identified predictable patterns in implied volatility forecasting using models like regression and VAR but found negative returns when applying these forecasts to VIX futures trading. Their best model was a simple linear regression with

seven economic variables, underlining the role of economic factors.

Paye (2012) explored how macroeconomic uncertainty, stock returns, and credit conditions influence volatility, finding that these factors Granger-cause volatility but don't significantly improve out-of-sample performance. Fernandes *et al.* (2014) used HAR models with economic variables, finding minimal long-term effects of the term spread on VIX, while Degiannakis *et al.* (2018) showed non-parametric models like SSA-HW outperform parametric ones for short-term forecasts.

Our study expands on this literature by analyzing a broader set of economic variables, using distinct phases for training, validation, and implementation to ensure trackable out-of-sample results. We also explore the underlying sources of predictability, considering both model design and variable selection, and tailor our design for practical applications, particularly in testing the profitability of VIX predictability.

3. Research design and the adaptive machine learning framework

In our study, we develop an adaptive learning methodology tailored for predicting VIX signals, addressing the broader challenges of implementing machine learning (ML) in financial forecasting.

3.1. Research design

From a research design perspective, we explore several critical questions:

1. Forecasting Objective: We aim to predict the daily directional signal of the VIX for the following day.
2. Explanatory Variables: Our approach involves a comprehensive selection of variables, ensuring real-time availability and the absence of look-back bias. Bloomberg serves as our primary data source, and we have identified 278 features across 14 categories, as detailed in table 1 (refer to Online Appendix I for the full list). These variables are selected based on economic theories and existing studies.†

   Our modeling approach can be summarized by the following generatic model form.

$$\hat{y}_{t+h} = f(X_{1,t}, X_{2,t}, \ldots, X_{278,t}; \theta), \tag{1}$$

   where:

   - $\hat{y}_{t+h}$ is the forecasted value of the target variable at time $t + h$. In this paper,
     - $y_{t+h} = \begin{cases} \text{up}, & \text{If } \Delta\text{VIX}_{t+h} > 0 \\ \text{down}, & \text{If } \Delta\text{VIX}_{t+h} > 0 \end{cases}$
     - Where ΔVIX_{t+h} represents the change in the VIX from time t to $t + h$ and $h = 1$.
   - $\boldsymbol{X_t} = \{X_{1,t}, X_{2,t}, \ldots, X_{278,t}\}$ is the set of predictor variables at time t which includes the 278 features derived from 14 groups of underlying data. These features include lagged variables of both predictor and dependent/target data. See online appendix for a detailed list of variables.
   - f is the machine learning model function.
   - θ represents the parameters of the machine learning model, determined during training.
3. Algorithm Selection: We incorporate a diverse range of ML algorithms, including Naïve Bayes (NB), Logistic Regression (LR), classic ML techniques like Decision Tree (DT) and Random Forest (RF), as well as more advanced methods such as Adaptive Boosting (AB), Multi-Layer Perceptron (MLP), and an Ensemble model (Ens) that integrates all the aforementioned algorithms. This selection covers a broad spectrum of model complexities to determine the most effective algorithm for predicting volatility direction. Further details on these algorithms are provided in Online Appendix II.
4. Hyperparameter Selection/Model Tuning: We delve into the specifics of each algorithm, including the selection of the most suitable model setup for our out-of-sample application. Additionally, we address how to systematically monitor and manage model performance decay during the inference process, focusing on retraining strategies.

We focus our discussion on the adaptive continuous learning methodology in the following.

3.2. The adaptive continuous learning methodology

Our methodology introduces an automated adaptive continuous ML framework. Figure 1 illustrates the essential elements of our closed-loop adaptive learning design, which encompasses three primary steps: training, validation, and implementation.

3.2.1. Step 1: training and model selection with dynamic hyperparameter setting and K-fold cross-validation. A key aspect of model construction is selecting the right hyperparameters for classification models. Traditional methods, often based on trial and error, are time-consuming and lack traceability. To address this, we use an AutoML-based Hyperparameter Optimization (HPO) method with Grid Search, combined with K-Fold cross-validation. This automates the selection process, identifying the optimal hyperparameters for each algorithm.

Given the challenges of time series data like the VIX, where maintaining temporal sequence is crucial, traditional cross-validation methods can introduce bias. We adopt a strategy that preserves chronological order, as suggested by Bergmeir *et al.* (2018), excluding entire rows for the test set to retain the time-dependent structure. Our approach ensures minimal information loss, particularly for time series features like MA5 and MA30. Following Bergmeir *et al.*, cross-validation remains valid if error terms are uncorrelated, which is likely

† There is a potential mixed frequency issue in the data. We take a snapshot of each point in time for all frequency of data at that point to construct input data for the forecast. We leave the determinant of usefulness of the feature by the algorithms instead of pre-modeling feature engineering.

Table 1. Summary of variables by groups.

Type	Variable descriptions	Num	Economic justification
Commodity	Bloomberg Commodity Index, WTI Crude Future, Brent Crude Future, Copper Future, Gold 100 Oz Future, etc.	9	Reflect market sentiment and economic conditions; e.g. oil prices influenced by geopolitical events.
Currency	Euro Spot, Japanese Yen Spot, British Pound Spot, Australian Dollar Spot, China Renminbi Spot, etc.	7	Indicate changes in economic conditions and investor sentiment across different regions.
Govt & Corp Bond	US Govt bonds (2, 5, 3 Yr, 12, 3, 6 Mth, 30 Yr), TII bonds, Corporate bonds, etc.	14	Reflect interest rate expectations and economic outlook; yield curve as predictor of economic activity.
Macroeconomic	ISM PMI, Unemployment Rate, PPI, Retail & Food Service, GDP, Labor Productivity, CPI, Consumer Confidence, etc.	57	Provide insights into overall economic health and consumer confidence, critical for predicting market volatility.
Major Equities	IBM, Apple, Amazon, General Electric, Microsoft, Bristol-Myers Squibb, FedEx, Nvidia, etc.	12	Represent significant market sectors; high-profile companies as market movers.
Seasonality	Day of the month, Day of the week, Week of the year, Month of the year, Day to next expired Wednesday	5	Seasonality effects can influence market behavior; certain times are associated with higher trading volumes.
SPX Member Tech	Technical indicators for S&*P* 500 Index, Bollinger Bands, Moving Averages, MACD, RSI, New highs/lows, etc.	35	Technical analysis indicators predict future price movements based on historical price patterns.
SPX Options & Futures	Historical call implied volatility, Put/call volume ratios, Option volumes, Open interest, Futures volume, etc.	16	Reflect market expectations of future volatility and risk, providing direct insight into investor sentiment.
SPX Subindex	Industry-specific indices within S&P 500 (banks, retailing, automobiles, transportation, software, insurance, etc.)	31	Sector-specific performance highlights trends and risks in different parts of the economy.
SPX Tech	Volatility measures, RSI, ARMS index, Money flow, Dividend per share, Volume measures, Moving averages, etc.	39	Provide detailed insights into market trends and investor behavior, essential for forecasting volatility.
Vix Tech	VIX-related technical indicators, Moving averages, RSI, Max/min days, Price change percentages, etc.	20	Understanding the historical dynamics and technical patterns of the VIX itself is crucial for predicting future volatility.
World Equity Index	Global indices (Dow Jones, Nikkei, Euro Stoxx, DAX, NASDAQ, FTSE, MSCI, etc.)	18	Global market trends influence domestic market volatility; international developments affect investor sentiment.

with our large models. We define a matrix of hyperparameter ranges for each algorithm and use K-fold cross-validation to ensure optimal performance and reduce overfitting risks (Appendix I for details).

3.2.2. Step 2: algorithm selection with out-of-sample validation. After identifying the best model setup for each algorithm, we conduct out-of-sample validation tests. Comparing training and validation accuracy informs us about the relative performance and stability of different algorithms. The algorithm with the best validation performance is selected for the next implementation step. We also consider the variability between training and validation performances, as large variations may indicate tendencies for overfitting or underfitting.

3.2.3. Step 3: implementation and closed-loop continuous learning. Over time, predictive model performance can decline as market behaviors evolve. The typical response is to periodically build new models with fresh data. However, this approach, often based on human judgment, can lead to delayed or unnecessary model updates.

To standardize and improve this process, we designed a closed-loop continuous learning framework. When model performance falls below a predefined threshold (e.g. a 42.5% prediction error rate in our study), a new training process is automatically initiated. Additionally, a stabilization period (e.g. at least 120 days in our research) is set before retraining to gather sufficient data for reevaluation and to prevent too frequent model switches. This approach ensures timely model updates, maintaining overall quality and performance with minimal human intervention.

3.3. The VIX sample and sub samples

The VIX is a real-time index that measures market expectations of 30-day volatility, often called the 'fear index.' Calculated by the CBOE from S&P 500 options prices, the VIX rises during market stress and falls during stability, providing a measure of market sentiment. Economically, the VIX serves as a key risk management tool for investors, indicating fear during downturns and complacency during calm periods. It also influences investment strategies, prompting safer investments in high-volatility periods and encouraging equity investments during low-volatility phases. VIX spikes often correlate with financial crises or major economic events, offering insight into potential disruptions.

Figure 2 shows VIX data from 1995 to 2020, divided into three periods. The blue line represents the In-Sample

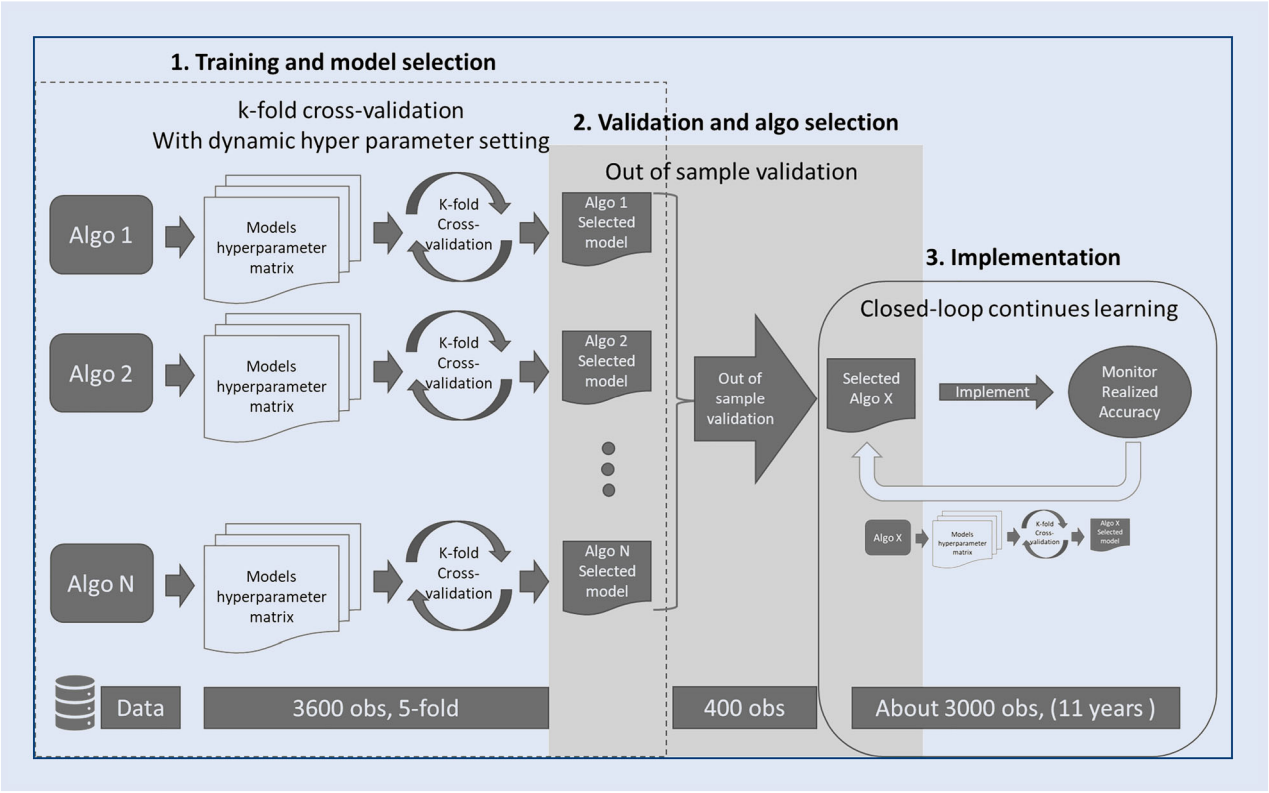

Figure 1. The adaptive continuous learning methodology.

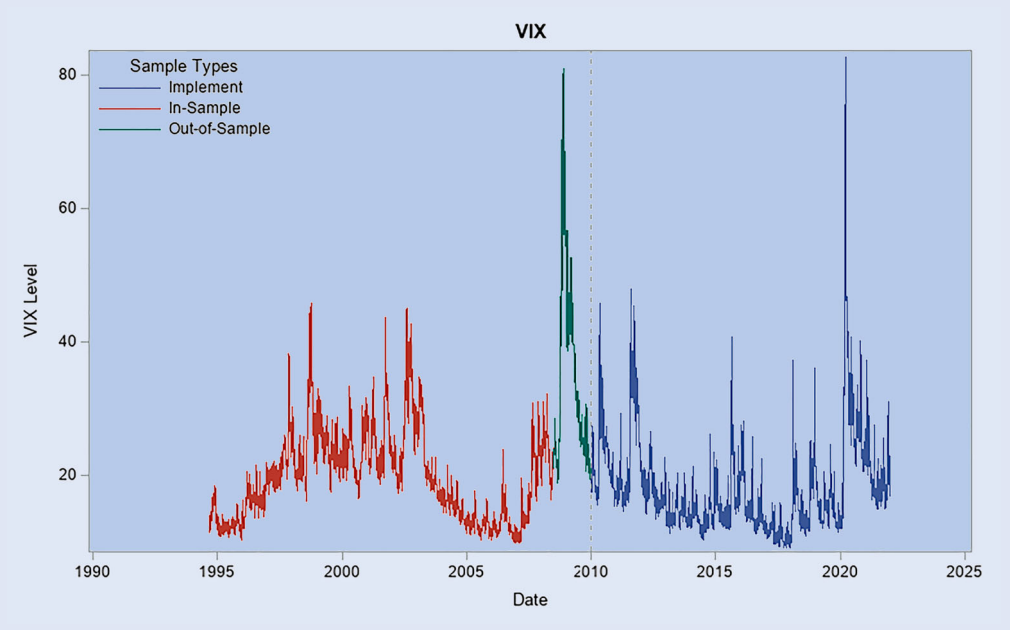

Figure 2. VIX time series plot. The chart visualizes the VIX sample from 1995 to 2020, segmented into three periods. The blue line (In-Sample Period) covers data up to 2009 for initial training and validation. The red line (Out-of-Sample Period) shows the final 400 observations of 2009 for validation and model selection. The green line (Implementation Period) spans 2010 to 2020 for testing the model. A vertical dashed line marks the end of 2009, separating the validation phases from the implementation stage.

Period (3600 observations up to 2009), the red line the Out-of-Sample Period (400 observations before 2010), and the green line the Implementation Period (2010–2020). A vertical dashed line marks the end of 2009, separating validation from implementation. During implementation, we use a rolling window of 3600 observations and 5-fold validation for retraining.

The VIX shows significant fluctuations, including clustering of high volatility, reflecting its strong connection to market stress. Notable spikes during the dot-com bubble, 2008 financial crisis, and COVID-19 pandemic underscore its role as a market sentiment gauge. The VIX also exhibits mean reversion, returning to its long-term average after short-term deviations. Summary statistics and dynamic properties

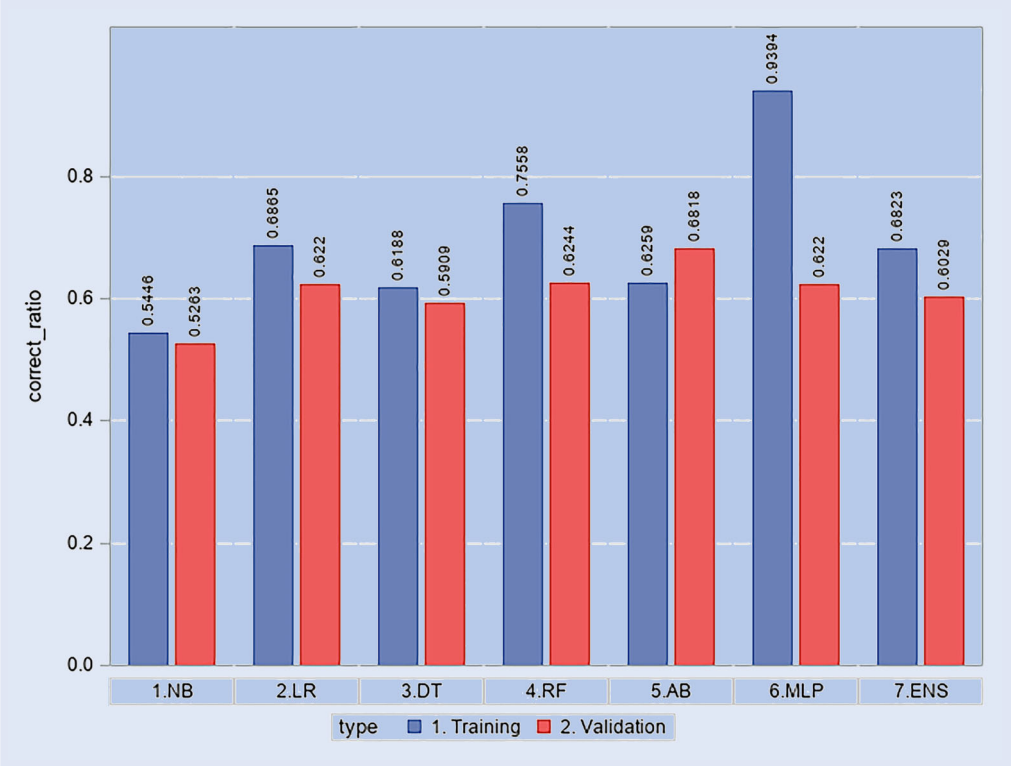

Figure 3. Training and validation accuracy for modelling at the end of 2009. This figure reports the accuracy ratios for the training and validation of each model.

across periods are discussed in Appendix II, showing close alignment between the training and implementation phases, confirming the model captures market volatility effectively.

4. Forecasting performance

Our empirical results are organized into three sections: prediction accuracy, economic evaluation, and sources of predictability. This section examines training, validation, and forecasting performance based on accuracy and market timing. Section 5 assesses the economic significance of the forecasts, while Section 6 explores the sources of predictability through various experiments.

4.1. Training and validation accuracy for modeling at the end of 2009

In evaluating the prediction accuracy of our selected models for all algorithms as of the end of 2009, we focused on the outputs from the 'Step 2 validation and algorithm selection.' After completing the K-fold cross-validation training, we retained the best-performing models and applied them to 400 validation data points. Figure 3 presents the training and validation accuracy for each algorithm, revealing four key observations.

Firstly, the Naïve Bayes (NB) model exhibited the lowest accuracy, indicating that more complex algorithms have a distinct advantage in this application. Secondly, a linear model like Logistic Regression (LR) demonstrated reasonable accuracy, suggesting that the predictability is significantly influenced by the economic relevance of the features we selected.

Thirdly, we observed a trade-off between model complexity and stability/variability, particularly when comparing in-sample training performance with out-of-sample validation performance. A notable decline in out-of-sample performance often signals overfitting. In this context, the Multi-Layer Perceptron (MLP), a neural network-type model with a complex nonlinear structure, showed signs of overfitting, evidenced by its high in-sample accuracy of 94%. However, its out-of-sample validation accuracy of 62.2% was still commendable, especially when compared to NB.

Finally, Adaptive Boosting (AB) yielded the best validation results. Intriguingly, its validation accuracy surpassed its in-sample accuracy, making it a standout choice for implementation in Step 3, following our framework's guidelines as of the end of 2009.

4.2. Out-of-sample implementation accuracy from 2010 to 2020

In this section, we present the yearly accuracy ratios for our out-of-sample forecasts by algorithms, as depicted in figure 4, with mean ratios detailed in table 2. Aligning with our validation results, figure 4 indicates that Naïve Bayes (NB) consistently shows the lowest accuracy rates. In contrast, Decision Tree (DT), Random Forest (RF), and AdaBoost (AB) maintain higher accuracy rates, consistently above 50%. The Multi-Layer Perceptron (MLP) underperforms, reinforcing concerns about overfitting identified during the validation stage. The Ensemble model (ENS) displays intermediate

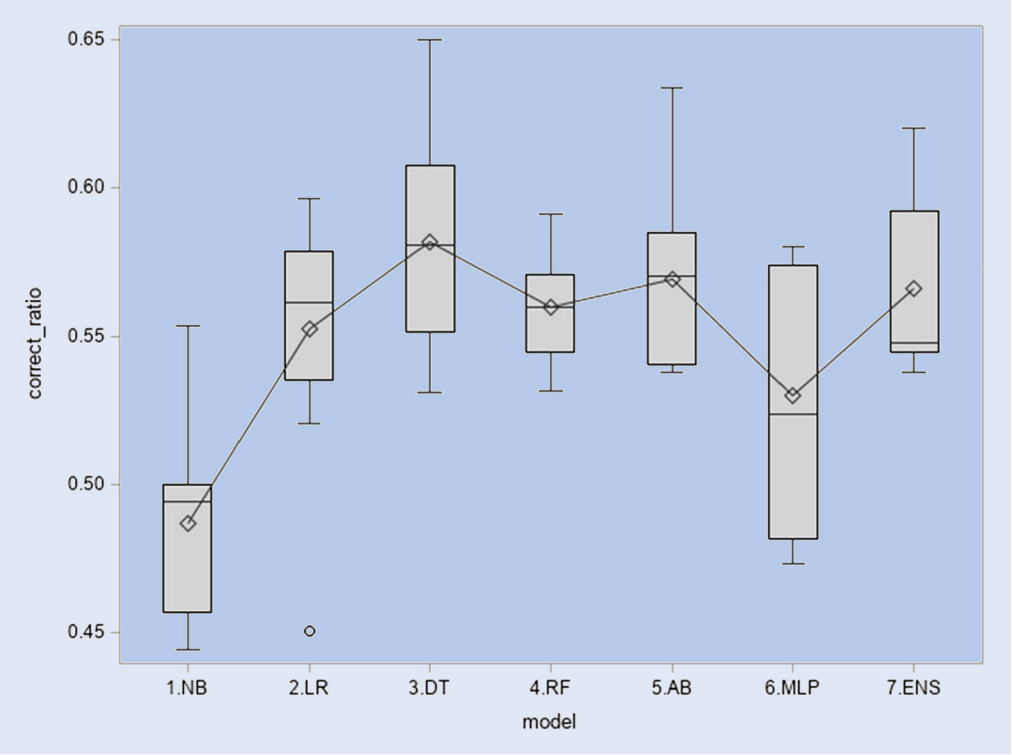

Figure 4. Out-of-sample implementation accuracy by model. This figure reports the box plot of the yearly correct ratio by model. The correct ratio is obtained from the out-of-sample forecast from 2010 to 2020.

performance between DT and RF and appears to reduce year-to-year variability in model performance, as evidenced by a narrower interquartile range.

Figure 5 reports the cumulative correct predictions for each model to demonstrate the dynamic of the implimentation accuracy. It can be interpreted as the number of net correct predictions at any given point in time. These plots demonstrate that decision tree family models, such as Decision Tree (DT), Random Forest (RF), and AdaBoost (AB), exhibit consistent growth in correct predictions, indicating stable and reliable performance over time. Conversely, more complex models, such as Multi-Layer Perceptron (MLP), show more varied performance across the time period analyzed.

Overall, these results affirm that both the simplest probabilistic classifiers (NB) and the most complex method (MLP) are less effective for directional forecasting, while decision tree-based models emerge as the most suitable for this classification task.

Regarding the model retraining during the closed-loop learning in Step 3, we assume the selection of the best models from each algorithm for implementation. By design, weaker algorithms in terms of performance necessitate more frequent retraining. Figure 6 shows that NB required the most retraining, with 23 instances over 11 years. Given our minimum run requirement of 120 days for each model, this frequency suggests that NB's performance consistently failed to meet our error rate threshold of 42.5%, indicating that retraining does not necessarily rectify performance issues in weaker algorithms. Conversely, DT and AB experienced less frequent retraining, approximately once a year. The Ensemble model demonstrated the most stability, requiring the least retraining (7 times in 11 years).

In terms of performance variation, despite frequent retraining, NB's performance showed low variability but consistently poor results. RF and MLP, on the other hand, exhibited significant variations in their training performance across different stages/models. The training and validation outcomes for all models align with our findings at the end of 2009 (figure 3), suggesting that our training regime consistently yields reliable outcomes across different datasets. Notably, the overfitting issue with MLP persisted throughout the closed-loop training.†

4.3. Statistical test

In the out-of-sample implementation phase, table 2 presents the accuracy and timing measures. To assess the differences in accuracy rates between models, we employ the Diebold and Mariano's (1995, DM) tests. Recognizing that the DM test may frequently reject the null hypothesis in small samples, Harvey *et al.* (1997, HLN) suggested modified statistics to mitigate this issue. Our main results include these HLN statistics.

To contextualize our findings within existing literature, we compare our results with a simple linear forecasting model, the Heterogeneous Autoregressive (HAR) model, known for its effectiveness in volatility forecasting as discussed in

† In the early stages of our study, we incorporated the Support Vector Machine (SVM) among our selection of models. However, we observed that this algorithm predominantly yielded one-sided predictions, demonstrating limited timing ability. To provide a comprehensive view, we have detailed the results and a focused discussion on the performance of SVM in an online appendix.

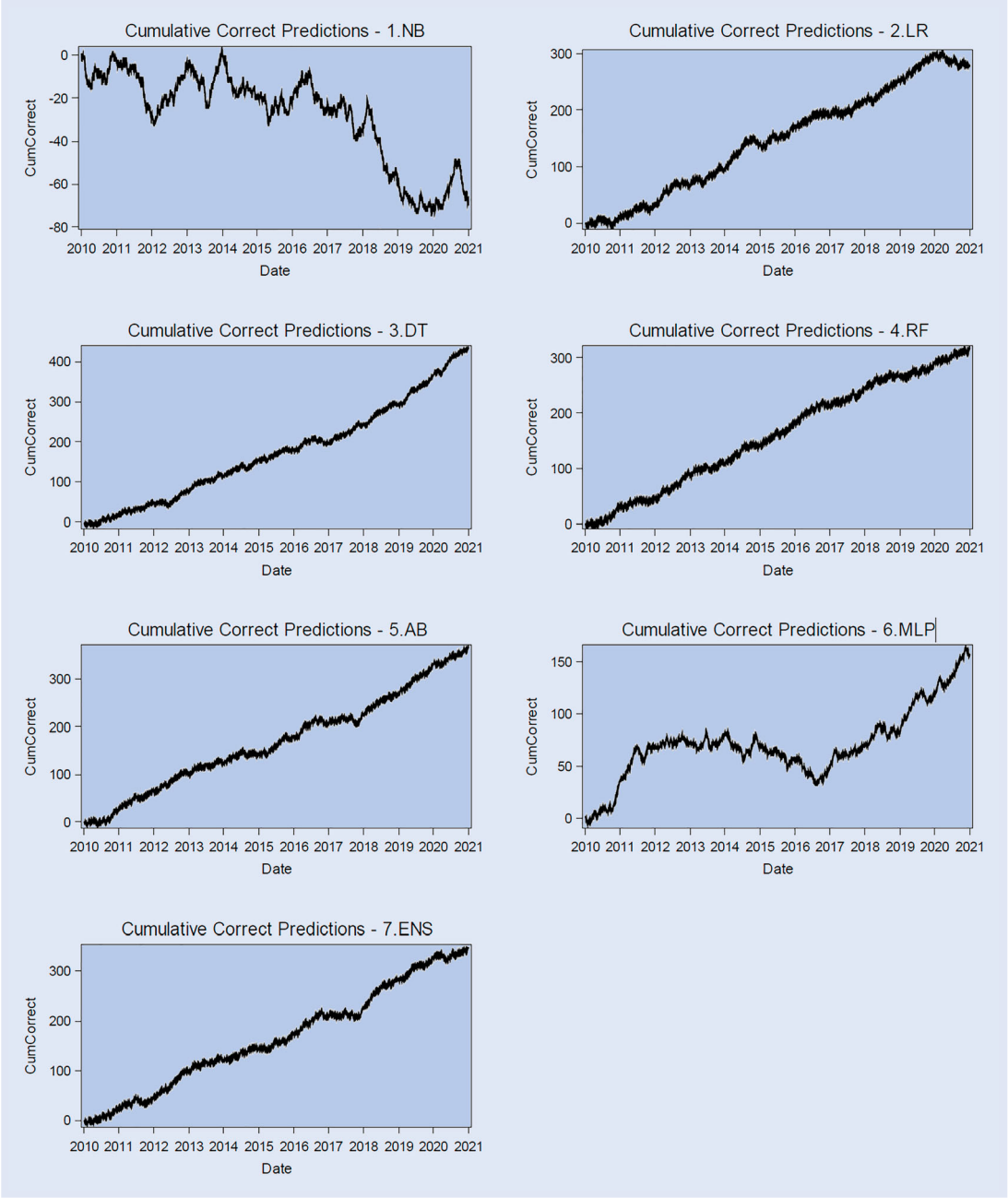

Figure 5. Implementation accuracy dynamics. This figure reports the cumulative correct predictions for each model. It can be interpreted as the number of net correct predictions at any given point in time.

Section 2. We implement a rolling daily HAR model forecast using three variables: lagged one-day, weekly average, and monthly average values of VIX, with a rolling window of 4000 observations, mirroring our main analysis.

Table 2 reveals several key findings. Firstly, Panel A confirms that all models, except NB and MLP, outperform the HAR model. The notable difference between HAR and Logistic Regression (LR), both linear models, lies in the number of features used. LR's outperformance over HAR underscores the significance of the additional economic variables included in our study for enhancing forecasting performance. Among the models, Decision Tree (DT) and AdaBoost (AB) stand out in accuracy, corroborated by information coefficients. For market timing, DT, AB, Random Forest (RF), and Ensemble (ENS) models all demonstrate over 10% market timing with robust statistical significance, whereas NB, MLP, and HAR show low market timing abilities.

Secondly, Panel B presents pairwise accuracy tests, indicating that DT outperforms all other models in accuracy. The Ensemble model surpasses only NB and MLP models.

In summary, the decision tree family models, including DT, RF, and AB, exhibit the highest prediction accuracy. These models consistently outperform the HAR model in predicting the directional movement of the VIX, highlighting their effectiveness in volatility forecasting.

Table 2. Forecast accuracy summary.

Panel A. Accuracy Rate						
	Correct ratio	Compared to HAR	Information coefficient		Timing ratio	
Models	Mean	HLN	Mean	*t*	Mean	*t*
1.NB	0.487	−3.91***	−0.0262	−1.36	0.020	1.35
2.LR	0.552	2.13**	0.1046	4.19***	0.097	4.94***
3.DT	0.582	4.45***	0.1634	6.93***	0.126	5.25***
4.RF	0.560	3.66***	0.1194	10.64***	0.122	9.60***
5.AB	0.569	4.12***	0.1383	8.01***	0.122	7.38***
6.MLP	0.530	0.39	0.0595	2.33**	0.057	2.61**
7.ENS	0.566	3.84***	0.1319	7.19***	0.119	6.41***
8.HAR	0.525		0.0502	2.31**	0.072	4.02***

Panel B. Pairwise HLN test on the accuracy rate difference between the row and column models							
	2.LR	3.DT	4.RF	5.AB	6.MLP	7.ENS	8.HAR
1.NB	−0.065***	−0.095***	−0.073***	−0.082***	−0.043***	−0.079***	−0.038***
2.LR		−0.030**	−0.008	−0.017	0.022	−0.014	0.027**
3.DT			0.022**	0.013	0.052***	0.016	0.057***
4.RF				−0.009	0.030**	−0.006	0.035***
5.AB					0.039***	0.003	0.044***
6.MLP						−0.036***	0.005
7.ENS							0.041***

Note: Panel A reports the correct ratio, information coefficient and timing ratio in the 11-year implementation period. The information coefficient is calculated by (2 × Correct_ratio) − 1. The timing ratio is calculated as (true positive ratio + true negative ratio) − 1. The HLN column reports the Harvey *et al.* (1997) test on the difference in accuracy rate between the model on the left and the HAR mode in the daily predictions of 11-year out-of-sample period. t-tests for the information coefficients and timing ratio are performed on the variations of the statistics among the 11 annual observations. Panel B reports the HLN test statistics on the difference in accuracy rate between the models in the rows and columns. ***, **, and * indicate statistical significance at 1%, 5% and 10% level respectively.

5. Economic evaluation: a simulated strategy

Diebold and Mariano (1995) stressed that the economic impact of forecast errors is context-dependent, influencing decision-making. To assess the economic value of our forecasts, we simulate a long-short trading strategy based on daily VIX prediction signals at market close. Although VIX itself is not directly tradable, this simulation tests signal accuracy by size-weighting returns.

We evaluate out-of-sample returns (figure 7) across models over 11 years. Higher accuracy generally correlates with better returns, though exceptions exist, such as Decision Tree (DT) yielding lower returns than Random Forest (RF) and Adaptive Boosting (AB), despite superior accuracy (t-value: 3.91). Table 3 shows models like ENS, RF, and AB exhibit high Sharpe ratios, with AB having the lowest drawdown.

Most models generate over 50 basis points in daily returns, annualizing to 125%. However, actual returns are lower due to tracking errors and transaction costs from rebalancing. This economic analysis shows that accurate VIX forecasts are more effective when market movements are larger, making them valuable for derivative portfolios and aiding market makers in SPX and VIX futures strategies.†

† We demonstrate two more realistic investment tests with some tradable VIX derivatives, considering transaction costs in an online appendix.

Table 3. Out of sample simulated long-short strategy performance.

	Daily return			Yearly MDD	
model	Mean	*t*	Sharpe	Mean	*t*
1.NB	0.0029	2.79**	0.85	−60%	−4.39***
2.LR	0.0060	5.06***	1.54	−52%	−2.87**
3.DT	0.0056	4.30***	1.27	−64%	−2.57**
4.RF	0.0078	6.85***	2.05	−47%	−2.84**
5.AB	0.0090	5.80***	1.73	−30%	−3.31***
6.MLP	0.0045	3.91***	1.18	−50%	−3.24***
7.ENS	0.0074	7.42***	2.24	−89%	−1.92*
HAR	0.0053	3.89***	1.15	−56%	−3.63***

Note: This table reports the mean daily return, the Sharpe ratio, and the average maximum annual percentage drawdown (MDD). t-tests are performed on the variations of the statistics among the 11 annual observations. ***, **, and * indicate statistical significance at 1%, 5% and 10% level respectively.

6. Source of predictability

This section explores the sources of predictability through various experiments. We assess model performance during major volatility events (6.1) and analyze the impact of economic variables using variable importance (6.2). Additionally, we examine two key model aspects: closed-loop training (6.3) and balanced sampling (6.4). We also compare multi-category versus binary predictions (6.5) and analyze prediction persistence by evaluating delays in predictors (6.6).

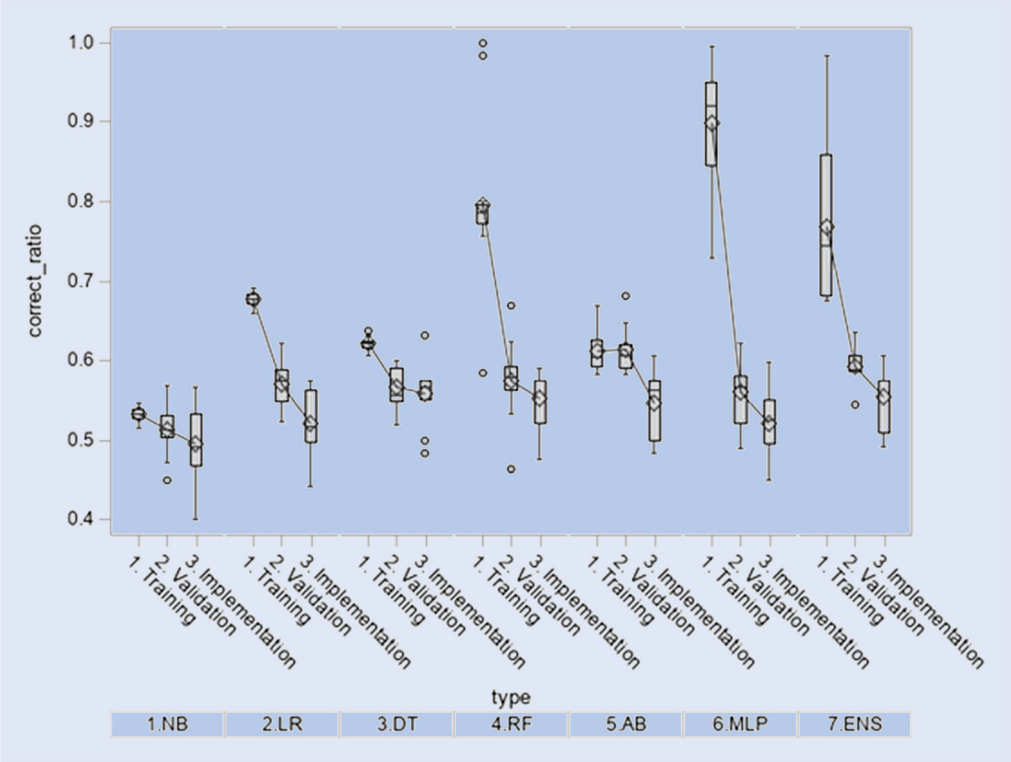

Figure 6. Retraining, validation, and implementation accuracy during the close-loop implementation period. This figure reports the distribution of accuracy for the models used in the implementation stage for each Algo. The numbers of training are reported at the bottom of the figure.

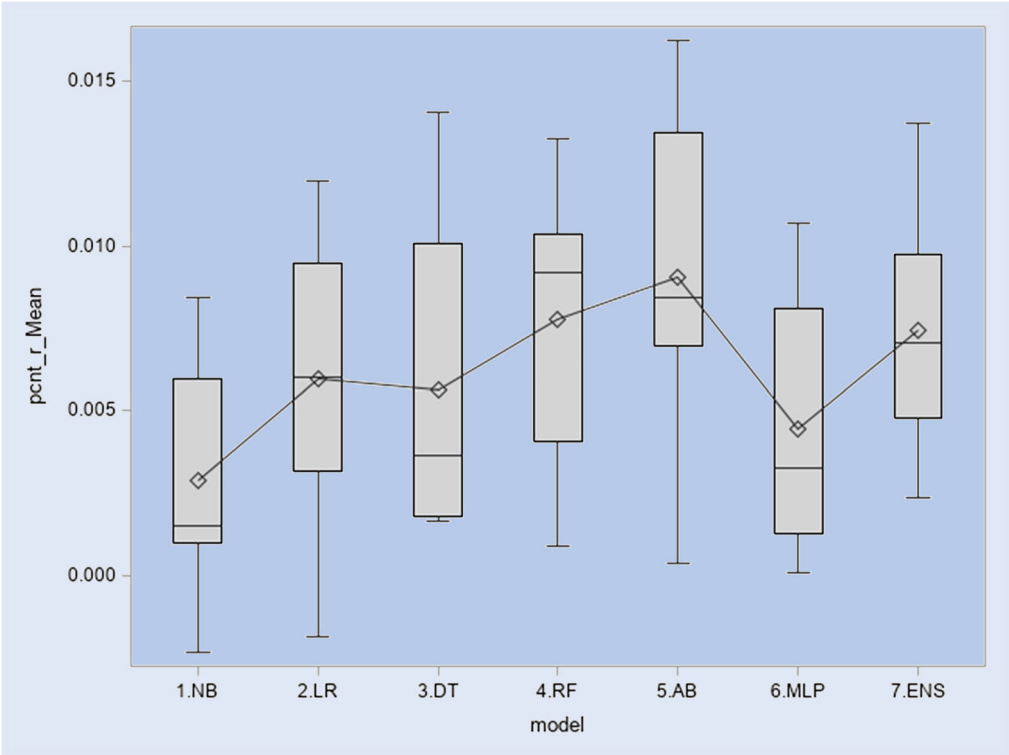

Figure 7. Out of sample simulated long-short strategy performance. This figure reports the distributions (box plots) of the mean daily return in the 11 years between 2010 and 2020 for each algorithm. The return is calculated by applying the predicted signal to the next day's VIX return. The diamond indicated the mean return.

6.1. Volatility spikes

Market volatility spikes, often caused by unforeseen events like the 2010 Flash Crash or the 2021 GameStop surge, create significant tail risks, particularly for shorting VIX. These spikes are largely exogenous and difficult to predict. We assess model performance around volatility spikes, defined as VIX changes over 20%. Table 4 shows a higher frequency of positive spikes (64) than negative ones (10). Models perform well during negative spikes with error rates under 20%, but struggle with positive spikes, where error rates often exceed 50%, except for Naïve Bayes (NB).

Table 5 tracks recovery from initial daily losses ($\sim$30%) after spikes. Decision-tree models, especially Adaptive Boosting (AB), show strong recovery, recouping 93% of losses within 20 days, compared to 23% for the HAR model. Most models fully recover within 60 days, except NB and MLP.

In conclusion, while VIX spikes are unpredictable, decision-tree ML algorithms demonstrate strong adaptability and recovery, highlighting their resilience in volatile conditions.

6.2. Variable importance and variable selections

Interpretable models in ML highlight the sources of predictability through variable importance metrics. We use the ExtraTreeClassifier, an ensemble method, to assess feature importance in VIX signal classification. This method calculates feature importance based on node impurity reduction, indicating a feature's influence on prediction accuracy.

Table 6 shows the top 20 variables by importance. The weekly jobless report consistently ranks highest across models, revealing its strong impact on market volatility.

Table 4. Prediction performance on VIX spikes days.

Models	Spikes	Number of days	Error days	% Err	Return		
					Mean	Min	Max
1.NB	negative	10	2	20%	0.1373	− 0.2591	0.2957
	positive	64	23	36%	0.0781	− 1.1560	0.5000
2.LR	negative	10	2	20%	0.1432	− 0.2337	0.2957
	positive	64	38	59%	− 0.0800	− 1.1560	0.5000
3.DT	negative	10	1	10%	0.1892	− 0.2327	0.2957
	positive	64	52	81%	− 0.2004	− 1.1560	0.4933
4.RF	negative	10	0	0%	0.2357	0.2050	0.2957
	positive	64	42	66%	− 0.1161	− 1.1560	0.4638
5.AB	negative	10	2	20%	0.1373	− 0.2591	0.2957
	positive	64	40	63%	− 0.0389	− 0.5000	1.1560
6.MLP	negative	10	4	40%	0.0449	− 0.2957	0.2696
	positive	64	32	50%	− 0.0037	− 1.1560	0.4933
7.ENS	negative	10	2	20%	0.1352	− 0.2696	0.2957
	positive	64	45	70%	− 0.1315	− 1.1560	0.5000
HAR	negative	10	3	30%	0.0958	− 0.2591	0.2957
	positive	64	40	63%	− 0.0884	− 1.1560	0.4933
SO	negative	10	0	0%	0.2357	0.2050	0.2957
	positive	64	64	100%	− 0.3126	− 1.1560	− 0.2022
SVM	negative	10	2	20%	0.1381	− 0.2591	0.2957
	positive	64	48	75%	− 0.1553	− 0.5000	1.1560

Note: This table reports the model forecasting performance for the negative and positive VIX spikes days. Spikes are defined as VIX movement greater or equal to 20%. It reports the number of days that spikes occur between 2010 and 2020. Error days (% err) reports the number (proportion) of days that the model makes incorrect predictions. The return columns report the mean, minimum and maximum in those spike days for each model.

Table 5. Return performance following the VIX spikes days.

Models	Initial losses	20 days after initial		60 days after initial		N
	on spike day (1)	Cumulated P&L (2)	Recover percentage (1)–(2)/(1)	Cumulated P&L (4)	Recover percentage (1)–(4)/(1)	
1.NB	− 0.3263	− 0.2709	3%	− 0.2566	15%	23
2.LR	− 0.3306	− 0.1538	46%	0.037	111%	38
3.DT	− 0.3157	− 0.1808	50%	0.1782	161%	52
4.RF	− 0.3266	− 0.1214	63%	0.1654	160%	42
5.AB	− 0.2812	− 0.0221	93%	0.2702	206%	40
6.MLP	− 0.3163	− 0.3526	− 6%	− 0.2178	32%	32
7.ENS	− 0.3158	− 0.119	62%	0.0793	135%	45
HAR	− 0.3208	− 0.2495	23%	0.0303	106%	40

Note: This table reports the mean initial losses for the incorrect predictions for the spike days. It reports the initial losses on spike day. It also reports the cumulated profit & losses 20 (60) after and including the spike day.

Table 6. Top 20 variable importance by ranking.

Panel A. Rank by Average of Variable Importance in the initial training

Rank	Name Full	Category	Rank in retrain
1	US Initial Jobless Claims SA change	Macroeconomic	1
2	day of the week	Seasonality	2
3	SPX Index pct members with new 52w highs	SPX Member Tech	28
4	SPX Index Volume	SPX Tech	45
5	S5TELS Index	SPX Subindex	64
6	VIX Index 60d	Vix Tech	132
7	SPX Index pct members with new 8w highs	SPX Member Tech	35
8	GBP Currency	Currency	53
9	S5AUCO Index	SPX Subindex	85
10	VIX Index RSI 14d	Vix Tech	27
11	VIX Index days diff min30	Vix Tech	15
12	SPX Index pct memb px blw lwr boll band	SPX Member Tech	144
13	S 1 COMB Comdty	Commodity	36
14	day of the month	Seasonality	4
15	SPX index days diff max30	Vix Tech	33
16	SPX Index volatility 260D	SPX Tech	112
17	SX5E Index	World Equity Index	103
18	US CPI Urban Consumers MoM SA	Macroeconomic	194
19	VIX Index RSI 30d	Vix Tech	9
20	S5INDU Index	SPX Subindex	183

Panel B Rank by Average of Variable Importance in the Retraining

Rank	Name Full	Category	Ranking in the initial training
1	US Initial Jobless Claims SA change	Macroeconomic	1
2	day of the week	Seasonality	2
3	SPX index RSI3d/RSI14d	SPX Tech	98
4	day of the month	Seasonality	14
5	VIX Index RSI 9d	Vix Tech	41
6	VIX Index RSI3d/RSI14d	Vix Tech	58
7	Day to maturity at next 3rd Wednesday	Seasonality	92
8	SPX Index RSI 3D	SPX Tech	24
9	VIX Index RSI 30d	Vix Tech	19
10	VIX Index RSI 3d	Vix Tech	112
11	CL1 COMB Comdty	Commodity	88
12	CO1 COMB Comdty	Commodity	50
13	SPX index days diff min30	SPX Tech	86
14	SPX Index RSI 30D	SPX Tech	85
15	VIX Index days diff min30	Vix Tech	11
16	SPX Index pct members with new 24w highs	SPX Member Tech	163
17	XAU Currency	Commodity	221
18	SPX Index pct members with new 12 wk lows	SPX Member Tech	31
19	VIX Index days diff max30	Vix Tech	48
20	SPX Index RSI 14D	SPX Tech	60

Note: This table reports the top 10 variables according to their ranking in each model and all models.

Seasonality variables like day of the week and days until VIX contract expiry also contribute significantly, reflecting patterns in investor behavior linked to the economic cycle. Technical indicators such as SPX's Relative Strength Index (RSI) and commodities like oil and gold are also influential.

While the top two variables remain consistent across models, other rankings show variability over time, with correlations between early and later models ranging from 63% to 77%. This highlights the need for regular model updates. Table 7 confirms that predictability stems from both technical indicators and fundamental factors like macroeconomic variables. Categories with individually small contributions, such as Macroeconomics, collectively provide significant insights. Notably, the VIX Techs group contributes only 11.46% of total importance, suggesting that focusing solely on VIX's historical data misses key explanatory variables.

6.3. *One-time model vs. dynamic continuous learning model*

To demonstrate the benefits of our continuous learning framework, we compared it to a one-time model approach where the model is built once and used throughout the 11-year out-of-sample period without updates. Table 8 shows that in the one-time model setup, only advanced models like Random Forest (RF), AdaBoost (AB), and Multi-Layer Perceptron

Table 7. Variable importance by category for all model summary.

Category	Mean	Min	Max	Sum	N
SPX Tech	0.0039	0.0031	0.0055	0.1981	51
Macroeconomic	0.0023	0.0003	0.0133	0.1432	61
SPX Subindex	0.0038	0.0033	0.0045	0.1238	33
Vix Tech	0.0041	0.0029	0.0054	0.1146	28
SPX Member Tech	0.0039	0.0032	0.0047	0.1003	26
World Equity Index	0.0039	0.0032	0.0044	0.0701	18
SPX Options and Futures	0.0039	0.0034	0.0045	0.0543	14
Govt & Corp Bond	0.0038	0.0032	0.0043	0.0527	14
Major Equities	0.004	0.0034	0.0045	0.0477	12
Commodity	0.0043	0.0038	0.0047	0.0391	9
Currency	0.0041	0.0038	0.0044	0.0284	7
Seasonality	0.0055	0.0039	0.0092	0.0276	5
All	0.0036	0.0003	0.0133	1	278

Note: This table reports the statistics for the average variable importance of all AB models used in the implementation stage including the one at the end of 2009. It reports the mean, minimum, maximum and sum variable importance and number of variables in each category. The rows in the table are ordered by the sum column. The conditional formatting with green is higher and red is lower in value within each column compared across different categories.

Table 8. Comparison between one-time model vs dynamic retrained model.

	Accuracy rate			
Models	RETRAIN1	ONEMODEL	Difference	HLN
1.NB	0.487	0.496	−0.009	−0.99***
2.LR	0.552	0.500	0.052	5.21***
3.DT	0.582	0.492	0.089	5.54***
4.RF	0.560	0.530	0.030	3.60***
5.AB	0.569	0.560	0.009	1.04
6.MLP	0.530	0.527	0.003	0.19
7.ENS	0.566	0.497	0.069	6.24***

Note: This table reports the accuracy rate in the 11-year implementation period for two different training approaches: one-time (Onemodel) and dynamic retrained (Retrain1). Onemodel uses the model trained at the end of 2009 and applies it to the 11 years without further retraining. Retrain1 is the methodology reported in the main results where retraining is triggered dynamically. The HLN column reports the Harvey *et al.* (1997) test on the difference in accuracy rate between the two training approaches. ***, **, and * indicate statistical significance at 1%, 5% and 10% level respectively.

(MLP) achieve forecast accuracy above 50%, resulting in modest gains from dynamic retraining.

The most significant improvements from dynamic learning are seen in the Decision Tree (DT) model, with accuracy rising from 49% to 58%, and the Ensemble model also benefits. However, frequent retraining worsens performance for Naïve Bayes (NB). These results highlight the importance of updating models with new data, particularly for decision tree-based models.

6.4. *Balanced vs unbalanced sampling*

A key concern with nonlinear models is the risk of one-sided predictions, where they perform well in training but poorly out-of-sample. To address this, we use a 'balanced' sample with equal up and down observations, selecting 4000 data points. This section compares the results of using balanced and unbalanced samples for the AdaBoost (AB) model.

Table 9 shows that while the unbalanced sample has a higher correction ratio and better information ratio, its timing ratio is significantly lower, indicating worse performance in predicting both up and down movements. Economically, the unbalanced model yields lower simulated returns (49 basis points vs. 90) and experiences a higher drawdown.

In conclusion, balanced samples improve the accuracy and robustness of predictions, reducing bias and enhancing economic performance across different market conditions.

6.5. *The size of VIX changes and multi-category forecasting*

A limitation of binary forecasting is that it doesn't account for the magnitude of market movements. High accuracy may lack economic significance if the model predicts minor changes correctly but errs during significant shifts. Our simulated investment strategy in Section 0 highlighted the importance of return-weighted signal quality. Here, we further explore the relationship between directional predictions and VIX movement size in two ways.

First, we introduce a 'size timing' measure, categorizing market changes into large and small based on the upper and lower 15th percentiles of past 250 observations. We assess accuracy in predicting big versus small changes. Second, we experiment with a multi-category prediction model (4D) with categories: up-small, up-big, down-small, and down-big.

Table 10 shows that the AdaBoost (AB) model's 4D approach doesn't improve overall accuracy compared to the binary model (2D). In fact, the market timing ratio is lower in the 4D model, and the binary model better captures large movements.

These results suggest no clear benefit from more granular forecasts, likely due to the 4D model's need for larger sample sizes, which our dataset cannot currently support.

Table 9. Results for the unbalanced sample of the adaptive boosting model.

Panel A. Accuracy					
	Correct ratio	Information coefficient		Timing ratio	
Training	Mean	Mean	*p*-value	Mean	*p*-value
Balanced	0.5691	0.1383	< .01	0.1224	< .01
Unbalanced	0.5714	0.1428	< .01	0.0895	< .01
Panel B. Simulated before cost return					
	pcnt_r_Mean			Yearly MDD	
Training	Mean	*p*-value		Mean	*p*-value
Balanced	0.0090	< .01		− 0.30	< .01
Unbalanced	0.0049	0.02		− 0.58	0.01

Note: This table reports the correct ratio, information coefficient and timing ratio in the 11-year implementation period for two different training approaches for the Adaptive Boosting model (AB). One uses a 'balanced' sampling approach which consists of equal amounts of ups and downs which is the same as the main result reported in table 1. The other uses an 'unbalanced' sampling approach simply taking 4000 data points at the time of estimation. The information coefficient is calculated by (2 × Correct ratio) − 1. The timing ratio is calculated as (true positive ratio + true negative ratio) − 1. *p*-values are from the test for the mean to be different from zero.

Table 10. Results of size-timing and multi-category prediction of the adaptive boosting model.

	Correct ratio	Information coefficient		Timing ratio		Big Correct ratio		Small Correct ratio	
NumD	Mean	Mean	*p*-value	Mean	*p*-value	Mean	*p*-value	Mean	*p*-value
2D	0.5691	0.1383	< .01	0.1224	< .01	0.6101	< .01	0.5522	< .01
4D	0.5683	0.1366	< .01	0.0871	< .01	0.5619	< .01	0.5731	< .01

Note: This table reports the correct ratio, information coefficient and timing ratio in the 11-year implementation period for two different training approaches for the Adaptive Boosting models (AB). One trains the model to predict up and down (2D) which consists of equal amounts of ups and downs which is the same as the main result reported in table 1. The other trains the model to predict four categories of movements up-small, up-big, down-small, and down-big (4D). The information coefficient is calculated by (2 × Correct_ratio) − 1. The timing ratio is calculated as (true positive ratio + true negative ratio) − 1. *p*-values are from the test for the mean to be different from zero.

Table 11. Persistence of prediction accuracy and the use of open-to-open predictions.

Panel A. Accuracy						
		Correct ratio	Information coefficient		Timing ratio	
Training	Application	Mean	Mean	t	Mean	t
C2C	C2C (main result)	0.569	0.138	8.01***	0.122	7.38***
C2C	C2C next day	0.541	0.082	4.49***	0.071	4.46***
C2C	O2O next day	0.527	0.054	2.82**	0.051	2.70**
O2O next day	O2O next day	0.565	0.130	4.35***	0.136	4.38***
Panel B. Simulated returns								
		Daily Return			Yearly MDD		
Training	Security	Mean	*t*	Sharpe	Mean	*t*	Min
C2C	**C2C (main result)**	0.0090	5.80***	1.73	− 30%	− 3.31***	− 87%
C2C	**C2C next day**	0.0020	1.38	0.43	− 69%	− 6.43***	− 142%
C2C	**O2O**	0.0021	1.50	0.46	− 51%	− 4.41***	− 93%
O2O	**O2O**	0.0069	3.69***	1.06	− 39%	− 3.93***	− 107%

Note: This table reports the accuracy and simulated returns for three different experiments with different training and application targets. C2C indicates the current close to the next period close VIX changes; O2O next day indicates the next day's open to the day after the next's open VIX changes. The training column reports the type of returns used to construct the predicted target while the application column reports the type of return used to calculate forecasting performance. Panel A reports the correct ratio, information coefficient and timing ratio in the 11-year implementation period for the different training and application approaches for the Adaptive Boosting models (AB). The information coefficient is calculated by (2 × Correct_ratio) − 1. The timing ratio is calculated as (true positive ratio + true negative ratio) − 1. Panel B reports the mean daily return, the Sharpe ratio, and the average maximum annual percentage drawdown (MDD). ***, **, and * indicate statistical significance at 1%, 5% and 10% level respectively.

6.6. Persistence of the prediction

This section examines the persistence of our model's predictions by addressing two key points: the impact of data delays (such as the 15-minute lag in equity data) and the stability of predictions over time. We assess the effect of delayed signals and explore the model's reliance on recent data for short-term forecasting.

We conduct three tests:

- **One-Day Delay Accuracy**: Assessing the impact of a one-day delay in applying predictions.
- **Next Day's Open-to-Open Accuracy**: Evaluating the prediction accuracy for the next day's open-to-open VIX movement.
- **Direct Open-to-Open Prediction**: Predicting the next day's open-to-open VIX movement using current day's close data.

Table 11 shows that applying a one-day delay reduces accuracy by about 3%, with notable declines in information and timing ratios, resulting in a 70 basis point drop in returns. Similarly, predicting the next day's open-to-open movement with the current day's close data shows slightly lower accuracy but still delivers 69 basis points daily.

These findings confirm that timely use of the signal is critical for maximizing forecast accuracy. For those constrained by data delays, directly predicting open-to-open movements remains a viable, albeit slightly less accurate, alternative. The results affirm the model's robustness, with better performance linked to more immediate data and consistent alignment between training and application targets.

7. Conclusions

Our study delves into the predictability of the VIX and identifies its underlying sources, demonstrating that daily VIX can be predicted with greater accuracy than existing literature suggests, and these predictions are economically significant. The predictability arises from two main areas: the crucial role of human input in selecting economically relevant variables and translating the prediction task into a format interpretable by machines, and the forecasting framework itself. Our extensive inclusion of economic and financial data, particularly the weekly jobless claim, contributes new insights to the literature on the determinants of market volatility and fear. This novel finding highlights an underexplored link between the labor market and overall market volatility, suggesting avenues for future theoretical and empirical research.

The development of our automated and adaptive training framework based on AutoML, focusing on explainability and trackability, addresses key challenges in applying ML to financial forecasting. This framework reduces human intervention, enhances algorithm selection and model tuning efficiency, ensures robust model validation, and includes proactive monitoring and retraining processes. These features collectively enhance the model's adaptability to market conditions and reduce the likelihood of biased predictions.

Furthermore, our study's findings have substantial implications for quantitative investment and risk management. The ability to accurately predict the VIX equips investors and risk managers with a vital tool for making informed decisions about asset allocation, hedging strategies, and risk exposure. The insights gained from key economic variables, especially the weekly jobless claims, provide a deeper understanding of market dynamics, enabling more sophisticated risk management approaches. The automated and adaptive ML framework developed in our study not only augments prediction accuracy but also adapts to evolving market conditions, ensuring robustness in investment and risk management strategies amidst market volatility.

In summary, our research contributes significantly to the fields of quantitative investment and risk management, offering advanced methodologies for predicting market volatility and enhancing the understanding of its determinants. This work paves the way for more effective and adaptive investment strategies in the financial industry, particularly in the face of unpredictable market movements.

Acknowledgements

We thank the associate editor and referee for their constructive advice and Guofu Zhou for his insightful comments on an early draft of this paper. We also thank Giuliano De Rossi for sharing his working paper with us in the early stages of the project. All errors are our own.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Supplemental data

Supplemental data for this article can be accessed online at https://doi.org/10.1080/14697688.2024.2439458.

ORCID

Yunfei Bai https://orcid.org/0009-0009-0270-6123
Charlie X. Cai https://orcid.org/0000-0003-1398-3715

References

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

Bergmeir, C., Hyndman, R. and Koo, B., A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Comput. Stat. Data An.*, 2018, **120**, 70–83. doi:10.1016/j.csda.2017.11.003.

BlackRock. Artificial intelligence and machine learning in asset management, 2019. Available online at: https://www.blackrock.com/corporate/literature/whitepaper/viewpoint-artificial-intelligence-machine-learning-asset-management-october-2019.pdf (accessed March 2021).

Bodie, Z., Kane, A. and Marcus, A.J., *Investment*, 2018 (McGraw Hill: New York).

Bollerslev, T., Generalized autoregressive conditional heteroskedasticity. *J. Econometrics*, 1986, **31**, 307–327.

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

Bucci, A., Realized volatility forecasting with neural networks. *J. Financ. Economet.*, 2020, **18**, 502–531.

CFA Institute. Artificial intelligence and machine learning in asset management, 2020. Available online at: https://www.cfainstitute.org/-/media/documents/book/rf-lit-review/2020/rflr-artificial-intelligence-in-asset-management.ashx (accessed January 2021).

Corsi, F., A simple approximate long-memory model of realized volatility. *J. Financ. Economet.*, 2009, **7**, 174–196.

Degiannakis, S., Filis, G. and Hassani, H., Forecasting global stock market implied volatility indices. *J. Empir. Financ.*, 2018, **46**, 111–129.

Diebold, F.X. and Mariano, R.S., Comparing predictive accuracy. *J. Bus. Econ. Stat.*, 1995, **13**, 253–63.

Donaldson, R.G. and Kamstra, M., An artificial neural network-GARCH model for international stock return volatility. *J. Empir. Financ.*, 1997, **4**, 17–46.

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

Erickson, N.,Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M. and Smola, A., AutoGluon-Tabular: Robust and accurate AutoML for structured data. Mar 2020. [arxiv: stat.ML] Available online at: https://arxiv.org/abs/2003.06505.

Fernandes, M., Medeiros, M.C. and Scharth, M., Modeling and predicting the CBOE market volatility index. *J. Bank. Financ.*, 2014, **40**, 1–10.

Hamid, S.A. and Iqbal, Z., Using neural networks for forecasting volatility of S&P 500 Index futures prices. *J. Bus. Res.*, 2004, **57**, 1116–1125.

Harvey, D.,Leybourne, S. and Newbold, P., Testing the equality of prediction mean squared errors. *Int. J. Forecasting*, 1997, **13**, 281–91.

Konstantinidi, E.,Skiadopoulos, G. and Tzagkaraki, E., Can the evolution of implied volatility be forecasted? Evidence from European and US implied volatility indices. *J. Bank. Financ.*, 2008, **32**, 2401–2411.

Kristjanpoller, W. and Minutolo, M.C., A hybrid volatility forecasting framework integrating GARCH, artificial neural network, technical analysis and principal components analysis. *Expert Syst. Appl.*, 2018, **109**, 1–11.

Maciel, L.,Gomide, F. and Ballini, R., Evolving Fuzzy-GARCH approach for financial volatility modeling and forecasting. *Comput. Econ.*, 2016, **48**, 379–398.

Paye, B.S., "Déjà vol": Predictive regressions for aggregate stock market volatility using macroeconomic variables. *J. Financ. Econ.*, 2012, **106**, 527–546.

Pedregosa, F.,Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É., Scikit-learn: Machine learning in python. [arxiv: cs.LG], June 2018. Available online at: https://arxiv.org/abs/1201.0490.

Psaradellis, I. and Sermpinis, G., Modelling and trading the U.S. implied volatility indices. Evidence from the VIX, VXN and VXD indices. *Int. J. Forecasting*, 2016, **32**, 1268–1283.

Whaley, R.E., The investor fear gauge: Explication of the CBOE VIX. *J. Portfolio Manage.*, 2000, **26**, 12–17.

Appendices

Appendix I. Detailing the utilized software, packages, and tuning parameters for replication

We built the ML framework using Keras, Scikit-learn, and AutoGluon, which are prominent open-source libraries for machine learning and AutoML (see Pedregosa *et al.* 2018, Erickson *et al.* 2020). Additionally, we extended these libraries by incorporating our unique closed-loop training and inference methodology, as described in Section 3.1. The parameters and their respective ranges used during the closed-loop AutoML training are provided in the following table.

Appendix II. Summary statistics and dynamic properties of VIX

This table reports the summary statistics and dynamic properties of the VIX across different sample periods: In-Sample, Out-of-Sample, Implement, and Full Sample.

The combined training sample (In-Sample + Out-of-Sample) and the Implement period exhibit similar overall statistics and dynamics, despite greater variation in the Out-of-Sample period. The mean VIX values are 19.83 (In-Sample) and 34.71 (Out-of-Sample), with the Out-of-Sample period's higher standard deviation (13.88) reflecting financial crisis volatility. The Implement period has a mean of 18.11 and a standard deviation of 7.18, with slightly higher skewness and kurtosis, indicating more extreme values. PACF values at lag 1 are high across all periods (~ 0.97), with diminishing correlations at higher lags. Overall, the training sample effectively captures market volatility characteristics

Tuning parameters			
Algorithm	Parameter	Selection range	Description
Logistic Regression	regularization strength	0.1–100	The strength of regulation
Decision Tree	max_depth	3–12	The maximum depth of the tree
	min_samples_leaf	8–16	The minimum number of samples required to be at a leaf node
	max_leaf_nodes	5–100	The maximum nodes a tree grows
Random Forest	n_estimators	1–50	The number of trees in the forest
	max_depth	1–100	The maximum depth of the tree
AdaBoost	n_estimators	1–100	The maximum number of estimators the boosting terminates
	learning_rate	0.001–1	Learning rate
XGBoost	n_estimators	50–150	The maximum number of estimators the boosting terminates
	max_depth	3–10	The maximum depth of the tree
	min_child_weight	3–10	The minimum sum of instance weight in a child
	learning_rate	0.01–0.1	Learning rate
MLP NN	alpha	0.01–1	Learning rate
	layer_1_size	1–100	The number of neurons in the first layer
	layer_2_size	1–50	The number of neurons in the second layer
	layer_3_size	1–20	The number of neurons in the third layer
Deep NN	batch_size	1024	Batch size for each learning
	epochs	50–75	The number of training iteration
	lambda	0.025–0.035	The regularization applied to the model.
	dropout_rate	0.10–0.25	Dropout rate
	learning rate	0	Learning rate
	layer_1_size	100–200	The number of neurons in the first layer
	layer_2_size	100–200	The number of neurons in the second layer
	layer_3_size	100–200	The number of neurons in the third layer
Ensemble	All parameters	All ranges	Combination of all the above parameters

Sample	Mean	Std Dev	Min	Max	Skewness	Kurtosis	PACF Lag 1	PACF Lag 5	PACF Lag 23	N
In-Sample	19.83	6.62	9.89	45.74	0.779	0.443	0.982	0.0597	0.0006	3600
Out-of-Sample	34.71	13.88	18.81	80.86	1.085	0.426	0.972	0.1174	0.0257	400
Implement	18.11	7.18	9.14	82.69	2.585	11.942	0.966	0.0608	− 0.0077	3131
Full Sample	19.91	8.32	9.14	82.69	2.128	7.769	0.979	0.0712	0.0083	7131