Predicting VIX with Adaptive Machine Learning

Yunfei Bai* and Charlie X. Cai**

Abstract

We used an automated machine learning framework to investigate economic factors that predict the CBOE implied volatility index (VIX), analyzing a comprehensive list of 278 variables for the first time. Our study tested multiple classification models, including Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, Adaptive Boosting, Multi-Layer Perceptron, and an Ensemble model that combined all methods. Based on the validation stage and an 11-year out-of-sample period, we found Adaptive Boosting to be the most effective classification model. We demonstrated the economic importance of this out-of-sample predictability by simulating a long-short strategy. We then focus on understanding the sources of predictability and the limitations of the method. Our tests revealed that both nonlinear methods and comprehensive economic variables were significant predictors of VIX, with the weekly US jobless report and some S&P 500 members' technical indicators emerging as the most important sources of predictability. While VIX spikes remained unpredictable, our algorithms could adapt quickly by extracting new information from the data to recover from losses in trading strategy tests. Our new evidence contributes to both machine learning applications in finance and practical volatility forecasting, offering alternative research designs and insights for traders and investors.

Keywords: Finance, Machine Learning, Volatility Forecasting, Quantitative

Trading, S&P500

GEL codes: G0, G17, C52, C55, C58,

* PhD, AI/ML and Big Data Consultant, Amazon Web Services Inc, US, felix.yunfeibai@gmail.com

** Corresponding Author, Professor of Finance, Liverpool University School of Management, University of Liverpool, Liverpool, UK. Email: x.cai7@liverpool.ac.uk. Website: www.CharliexCai.info

Acknowledgement:
We thank Guofu Zhou for his insightful comments on the early draft of this paper. We also thank Giuliano De Rossi for sharing their working paper with us in the early stage of the project. All errors are ours.

1 Introduction

The importance of the CBOE Volatility Index (VIX) cannot be overstated given its role as one of the key indicators for practitioners and policymakers to gauge the forward-looking market condition. Coined as the 'fear' index (Whaley, 2000), the predictability of this index has far-reaching consequences beyond the US financial market. However, the extensive attention to this index makes it one of the most efficient markets to incorporate news in the global financial market. This makes it hard to predict given the existing information which seems to be the case in existing studies[1]. However, most of the existing studies are only confined to a small number of predictors and/or use linear time series methods. This is partly due to the constraint of the linear method applied. The objectives of this paper are two folds: 1) to study to what extent, from both forecasting accuracy and economic importance point of view, the VIX can be predicted with a large number of economic indicators enabled by the recent development of Machine Learning (ML) methodology and 2) to understand the source of the predictability and the potential limitation of the approach.

The improvement in machine learning (ML) in recent years and the growth of computing power have enabled many new applications[2]. Most of the published and working papers in the finance-ML literature have mainly focused on cross-sectional return forecasting and lower frequency such as monthly or annually. Relatively little work has been done on forecasting the time series volatility[3]. We start our research with the objective to predict VIX's next day's directional movement. We choose to predict direction instead of the level of

[1] There is a line of liteature use VIX to proxy for uncertainty which is similar to the definition of unforecastable part of the economy(Jurado, Ludvigson, and Ng, 2015) and Bali and Zhou, 2016).

[2] For survey studies, see Blackrock's discussion on ML in asset management https://www.blackrock.com/corporate/literature/whitepaper/viewpoint-artificial-intelligence-machine-learning-asset-management-october-2019.pdf [accessed March 2021] and CFA institute's report in similar topic. https://www.cfainstitute.org/-/media/documents/book/rf-lit-review/2020/rflr-artificial-intelligence-in-asset-management.ashx [accessed Jan 2021]

[3] Most of the existing study in volatility forecasting either adopt a univariate approach such as GARCH or HAR type model or swith limited number of economic explanatory variable. See a brief review in section 2.

VIX as it matches with the operational objective of an investment decision which is ultimately a binary one (i.e., to long or short)[4].

For choosing candidate explanatory variables, this is the part that human (i.e., the 'domain experts') plays an important role. We try to be as comprehensive as possible while making sure the data is available in real-time without a look-back bias. To this end, we use Bloomberg as our main data source. Building on existing literature, we collected a unique list of 278 features in 14 categories. They include time-series history of the S&P 500 index (hereafter, SPX) information covering both fundamental and technical indicators, global market indexes, industry subindexes, major commodities, foreign exchange markets, corporate and government bonds, US macroeconomic indicators and, last but not least, seasonality (e.g., days of the week). The underlining source of information is motivated by both theories of fundamentals such as the change of economic conditions, changes of market liquidity, and the spillover of global shock as well as technical such as the momentum/reversal and calendar effects[5]. We start the data collection in 1993 when the VIX index is introduced. What has not been considered in the current models is the soft and unstructured information such as news sentiments partly due to the lack of sufficient historical data.

In terms of ML algorithms, we choose six algorithms including more traditional methods such as Naïve Bayes (NB), Logistic Regression (LR) and classic ML such as Decision Tree (DT) and Random Forest (RF); we also include more advanced methods such as Adaptive boosting (AB), Multi-Layer Perceptron (MLP) and an Ensemble model using all of the above[6].

---

[4] Although we choose to predict direction, in our empirical analyses when we make comparison among the competing models, we study the 'profit' of a strategy that following the prediction direction. This is effectively comparing the performance of the models taking into consideration of the magnitude.

[5] We process the data carefully to help the ML understand the 'context' of the data while avoiding using future information in our data transformation. For example, since the machine only see one observation and will not have the time series context of the observations, most of the variables entering into the system as changes to capture the new information in the variables. For key variables such as the SPX return and changes of VIX, more elaborated historical data points are included as individual features in the input data set. We also constructe a 'balanced' sample with equal amount of the realized up and down sample during the traing step. More discussion about this in later and in Section 6.3.

[6] In this paper, we used the term 'algorithm' to refer to a modelling methodology (such as Decision Tree) while the term 'model' is used to refer to a specific parameterization of the

After proposing our machine learning framework for this application, our empirical study is organized into three key sections with distinctive objectives of studying 1) forecasting performance, 2) economic evaluation and 3) source of predictability.

For **forecasting performance**, we have the following key findings. First, examining the training and validation accuracy of the 'best' models of the algorithms at the end of 2009, we find there is a trade-off between model complexity and stability/variability of the model when comparing the in-sample training and the out-sample validation performance[7]. NB has the lowest *training* accuracy (54.5%) while the neural network type model MLP has the highest (93.94%) suggesting more complex algos help to fit the data better (lower bias). However, there is a much larger drop in the out-sample validation in the MLP (to 62.2%) than NB (to 52.6%). This is clear evidence of overfitting for MLP (i.e., high variance)[8]. AB provides the best validation results of 68.2%. Interestingly, its validation results are higher than the in-sample results of 62.6% (see Figure 2). Should we make a decision at the end of 2009, AB would be the one we pick to take forward to implementation in the out-of-sample period. Second, the out-of-sample implementation produces a pattern that is broadly in line with that of the validation results suggesting that the validation plays a useful role in assessing the model's performance. The decision-tree type models produce the best performance.

We then turn to an evaluation of the **economic importance** of the model forecast. Forecasting daily VIX has many potential applications. For example, it can be used with other valuation and risk management models to forecast the directional changes of the next day's valuation and value at risk. More directly, it will inform the derivative market makers for VIX-related

---

algorithm. This distinction is only important in the discussion of training and tuning. In the discussion of the later part of the process, these two terms are used interchangeably.

[7] For the training sample, we use 4,000 data points for training which is about 16-year worth of data. In which 90% of the data is used for training with 5-fold cross-validation and 10% for out-of-sample validation. We start our out-of-sample implementation from Jan 2010 and end in Dec 2020. For testing purposes, instead of reporting only the choosing algorithm's out-of-sample in the implementation period. We will report all of the algorithms' performance in the closed-loop implementation.

[8] See James et al. (2013) Section 2.2.2 page 33- 36 for discussion about the bias and variance trade-off. Intuitivly, a model with high variance is the one's performance is more sensitive to the change of sample.

products about the potential market movements. To illustrate the economic relevance, we simulate an 'investment' strategy to long and short the VIX index[9]. We show that it produces an average daily return ranging from 6 (LR) to 90 (AB) basis points. For the AB model, this is annualized to 225% (without compounding) and a Sharpe ratio of 1.7 based on the yearly distribution of the average returns. This model also has the smallest average maximum annual drawdown (MDD) at 30% which measures the lowest cumulative return from the beginning of each year (Table 3). Since VIX itself is not directly tradable, this exercise can be seen as measuring the economic importance of the directional forecast by weighting the predictions with the size of the VIX movement. This evidence suggests that the directional forecast captures VIX changes with significant sizes.

One of the biggest demands when studying VIX predictability is about the unpredictable event i.e., volatility spikes such as the flash crash in 2010, Brexit referendum in 2016, and inflation in employment numbers in 2018. We show that these events (larger than 20% daily movement) are truly exogenous to the system. As there is no information from the data point that the machine can learn from our models cannot predict these spikes. Importantly, the models can adapt to the new information relatively quicker and recover a large part of the losses within 20 days and fully recover in 60 days. These performances are much better than the non-model short-only strategy and other models.

The ML model's performance also compares superiorly to existing studies. We dedicate a large part of our study to understanding **the source of predictability**. This is done through variable importance analyses and experiments with the model and sampling settings to attribute the main model's performance.

For variable importance (VI), we find that the variables are used in quite a similar way across different models. We examine the top 20 most important variables and show that the US weekly jobless report consistently plays the most

[9] We understand that VIX itself is not directly tradable. This exercise is used to quantify the magnitude of economic importance of the directional forecast instead of testing the profitability of the strategy. We will test the profitability of investment strategy on tradable instruments in an online appendix which produce consistent conclusions.

important role in all models. Seasonality variables such as day of the week, day of the month or number of days to the next VIX futures contract expiry also have relatively high contributions in the models. The findings in the top 20 variables confirm the technical nature of these short-term prediction exercises. The underlining source of predictability follows similar arguments for technical analysis. It is likely due to the reversal or momentum effect driven by the behavioral or liquidity condition of the market. Nevertheless, it also shows that the new information in the economic condition, such as employment data, can also predict short-term volatility dynamics. The channel of such impact could be potentially through the heterogeneity in interpreting the new economic news which extends the persistence of volatility to the next period. Importantly, the VIX Techs group contributed only 11.46% out of the 100% among the variables. This suggests that the existing studies using only VIX's historical data such as the HAR model will have missed many potential explanatory variables.

We also conduct a series of tests on the model specification. First, we find that dynamic retraining provides significant improvement to most models. Second, we show the benefit of using a 'balanced' training sample to reduce an unconditionally one-sided prediction and improve the accuracy for both directions' predictions. Third, we show that our models perform better in predicting big than small price movement. Intuitively this is consistent with the view that a bigger directional change may be relatively easier to detect than a small size change which may go either way due to noise. Fourth, we extend our original binary prediction into a four-category prediction model (4D): up-small, up-big, down-small, and down-big. The overall forecasting accuracy is the same as the original binary predictions. Fifth, we study the persistence of prediction. We show that applying the signal with some delay such as to the next day's open-to-open return or the next day's close-to-close would still produce higher than 50% accuracy for the AB models. The model performance decreases as there are gaps between the data and the predicting target. The automated ML structure can produce reasonable predictions, without modification, when the predicting objective is changed which provides evidence for the robustness of the model methodology.

Overall, we show that the model performance is driven by the combination of information embedded in the economic variables selected and the flexibility of the modelling structure to extract the nonlinear relationship from the data.

Our study makes two significant contributions. Firstly, we demonstrate the superiority of machine learning methods over traditional techniques, such as Logistic Regression and HAR, in forecasting volatility. Our work builds on the VIX forecasting studies of Konstantinidi, Skiadopoulos, and Tzagkaraki (2008), Paye (2012), and Fernandes, Medeiros, and Scharth (2014). Most machine learning applications in finance focus on capturing nonlinear relationships using a limited set of economic predictors. Our study, however, expands on this approach by incorporating a more extensive range of economic variables, highlighting the importance of expert input in the research design. This enabled us to comprehensively study one of the most critical benchmarks in the financial market. In particular, we identified the weekly US jobless report as an essential predictor of volatility, despite the lack of direct empirical evidence linking it to market volatility. This news item remains one of the most closely watched economic indicators on the Bloomberg system, and our study demonstrates the value of including it as a predictor of market volatility.

Secondly, we offer valuable insights for future research on machine learning in financial forecasting. Our proposed closed-loop adapted learning framework consists of AutoML and HPO for algorithm and hyperparameter selection, as well as performance-based closed-loop continuous learning. This framework reduces manual model setup, avoiding overturning parameters, and enables the continued out-of-sample application of a model with adaptive updates. Our empirical evidence demonstrates the advantages of this design, which could be useful for future studies on ML in financial forecasting.

Overall, our findings and methods provide valuable insights for investment and risk management practices, such as volatility timing trading strategies or managing portfolio risk exposure. Our study offers a new perspective on volatility forecasting, highlighting the advantages of machine learning techniques in the financial market.

The rest of the paper is organized as follows. Section 2 gives a brief discussion of the volatility forecasting literature. Section 3 describes our research design. Sections 4, 5 and 6 report the empirical results for forecasting performance, economic evaluation, and source of predictability. Section 7 concludes.

2 Related literature

2.1 Historical and realized volatility forecasting

The most classic volatility forecasting model is the GARCH family model (Engle, 1982 and Bollerslev, 1986). For direct modelling of volatility time series, Andersen, Bollerslev, Diebold, and Labys (2003) showed that direct modelling of multivariate realized volatility outperforms, in terms of out-of-sample forecasting, the popular GARCH and stochastic volatility models. In this line of development, the heterogeneous autoregressive (HAR) model by Corsi (2009) has become one of the popular benchmark models given its ability to capture the persistence in the volatility series. It is a simple AR-type model of the realized volatility with the feature of considering volatilities realized over different time horizons including one day, five (weekly) and 22 (monthly) days. Despite its simplicity, the HAR model proves to be able to reproduce the volatility persistence observed in the empirical data.

There is a considerable body of literature trying to combine GARCH or HAR with nonlinear or nonparametric methods to improve forecasting accuracy (Kristjanpoller and Minutolo (2018), Maciel, Gomide, and Ballini (2016), Psaradellis and Sermpinis (2016)). Donaldson and Kamstra (1997) show that an Artificial Neural Network -GARCH model is found to generally outperform its traditional competing models—GARCH, EGARCH, and Sign-GARCH models—in both the in-sample and out-of-sample period when studying the performance of stock return volatility forecasting models using daily returns data from London, New York, Tokyo, and Toronto. More recently, Bucci (2020) studied monthly realized volatility and shows that long short-term memory (LSTM) Recursive Neural Network (RNN) can outperform the linear

models in out-of-sample forecasts. Most of the studies in this area focus on testing and comparing the time series forecasting methods with no or a limited number of other economic determinants in the system.

2.2 Implied volatility and VIX forecasting

Implied volatility is a measure of future expectations embedded in the options price. It has gained its prominent place in the world since CEBO introduced the VIX index in 1993. The VIX Index is an established and globally recognized benchmark of U.S. equity market volatility also known as the fear index. It measures the 30-day expected volatility of the U.S. stock market, derived from real-time, mid-quote prices of SPX call and put options. Therefore, VIX itself can be seen as a forecast of future volatility. Early work by Hamid and Iqbal (2004) shows that using neural networks to forecast the volatility of S&P 500 Index futures prices can outperform implied volatility forecasts. However, the objective of our study is to forecast VIX itself. In this regard, Konstantinidi, Skiadopoulos and Tzagkaraki (2008) show that predictable patterns are detected when studying implied volatility forecast with the regression model, VAR, and principal component analysis. Especially, for VIX they find accuracy as high as 54.7% in the two and half years of out-of-sample study until September 2007. However, when they directly apply these signals to trade VIX futures. All returns are negative. Interestingly, out of all of the models, the best model is the simple linear regression model with seven economic variables such as interest rate, Euro/USD exchange rate, WTI (Brent Crude Oil); the changes of the 30-day historical volatility, the changes of the slope of the yield curve, and the change of futures contract volume. This suggests the importance of the economic variable in this forecasting exercise. Paye (2012) find that several variables related to macroeconomic uncertainty, time-varying expected stock returns, and credit conditions Granger cause volatility. However, he finds no evidence that forecasts exploiting macroeconomic variables outperform a univariate benchmark out-of-sample. Whether or not and which economic variables will be relevant to volatility forecast, especially for relatively high frequency such as daily, is still a research question.

More recently, Fernandes, Medeiros, and Scharth (2014) use both parametric and semiparametric heterogeneous autoregressive (HAR) models with additional economic variables to forecast VIX. Among the economic variables, they find that the term spread has a slightly negative long-run impact on the VIX index. Importantly, they show that it is pretty hard to beat the pure HAR process because of the very persistent nature of the VIX index. Degiannakis, Filis, and Hassani (2018) show that non-parametric models of Singular Spectrum Analysis combined with Holt-Winters (SSA-HW) for univariate forecasting have statistically superior predictive ability for short-term implied volatility forecasting compared to parametric models such as the pure HAR model[10].

Relative to these existing studies, our study is different in four ways. First, compared to the studies using univariate information or limited among of economic variables, we can study a much larger set of economic variables. Second, our research design provides clear training, validation, and implementation explicitly which makes our results trackable and the test in the implementation stage is close to true out of sample by design. Third, compared to most of the studies with more advanced nonlinear modelling methods, we focus not only on the testing of the model accuracy but also on studying the underlining source of predictability (from both the model design and variable choices). Finally, we design our experiment with practical application in mind with careful consideration of data delay and the type of instruments in our practical application tests. So far, there is limited evidence of the potential profitability originating from the predictability of VIX.

3 Research Design, Sample and Measurement

We design an adaptive learning methodology for the VIX signal prediction. This framework is developed to address some of the challenges of implementing

[10] However, non-parametric models are very flexible. Fitting the non-parametric model without a clear validation strategy could leads to overfitting despite the final results are prested as rolling window 'out-of-sample' foreasst.

ML in the financial forecast in general. From a researcher's point of view, we have the following questions:

1. What is the objective of the forecast (**forecasting target**)?
2. What is the relevant information that should be included as explanatory variables (**feature selections**)?
3. Which ML algorithm we should include in our analysis (**Algorithm selection**)?
4. What specification of a given algorithm we should choose (**hyperparameter selection/model tuning**)?
    a. Which algorithm and selected model setup we should choose to apply to our problem out of sample?
    b. How to monitor and address the model performance decay during the inference process systematically (**Retraining**)?

We address these questions in the following. The objective of this research work is to predict the VIX daily signal for the next day. In selecting the variables, we try to be as comprehensive as possible while making sure the data is available in real-time without a look-back bias. To this end, we use Bloomberg as our main data source. Table 1 summarizes the 278 features in the 14 categories (See Online Appendix I for the full list). These variables are broadly informed by economic theories in existing studies[11].

<Insert Table 1>

For the main analyses of this paper, we include Naïve Bayes (1.NB), Logistic Regression (2.LR), and classic ML such as Decision Tree (3.DT) and Random Forest (4.RF); we also include more advanced methods such as Adaptive boosting (5.AB), Multi-Layer Perceptron (6.MLP) and an Ensemble model (7.Ens) using all of the above. These cover a wide range of model complexity to

[11] 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 engerneering.

examine which type of algorithm is better for volatility directional prediction. Further details of these algorithms are in Online Appendix II. We focus our discussion of the adaptive continuous learning methodology on the following.

3.1 The adaptive continuous learning methodology

What is new in our methodology is the automated adaptive continuous ML framework. Figure 1 summarizes the key elements of our closed-loop adaptive learning design which consists of three key steps: training, validation, and implementation.

<Insert Figure 1>

3.1.1 Step 1 Training and model selection with dynamic hyperparameter setting and k-fold cross-validation

Selecting the right hyperparameters for the classification model is an important step in the process of model construction and tuning. The normal approach is through trial and error. It depends on human experience and can be time-consuming and potentially untraceable. In this research, we employed an AutoML-based Hyperparameter Optimization (HPO) method with Grid Search. A K-Fold cross-validation method is used to automate the selection process and generate the best set of hyperparameters for each algorithm of the classification model.

In this tuning technique, we build a matrix of pre-defined ranges of the hyperparameters for each algorithm. Once all the combinations are evaluated, the model with the set of parameters that give the top performance is considered to be the best. The training for each set of the parameter is through K-fold cross-validation which utilizes a data set randomly partitioned into K mutually exclusive subsets. Out of the K sets, one is kept for testing while others are used for training. Throughout the whole K folds, the training process is iterated to achieve optimal convergence and avoid the possibility of over-fitting to one fold of the data.

3.1.2 Step 2 Algorithm selection with out-of-sample validation

Once the best model setup of each algorithm is trained and identified. We subject these 'best' models of different algorithms to another round of out-of-sample validation tests. Comparing the training accuracy and the validation accuracy would further inform us about the relative performance and variation of different algorithms. We can then choose the algorithm that has the best validation performance as the main model and proceed to the next step of implementation of the model. In addition to taking into consideration of forecasting accuracy during the valuation stage, the variability of performance between training and validation will also be considered. Models with large variations between these two stages may indicate the tendency of the models to be over- or under- fitting. A low variation is preferred.

3.1.3 Step 3 Implementation and closed-loop continuous learning

The predictive model performance can drop with time, as the market can have new behaviors that were not captured by the model using relatively old data samples when training. The typical approach to address this issue is to collect new data samples regularly to build a new model to replace the old one. However, the replacement cycle is determined based on human experience, therefore the model switch could be either too late or unnecessary.

To further standardize this process, we designed a closed-loop continuous learning framework. When the model performance starts to drop below the target (for example, in this paper we set the prediction error rate at 42.5%)[12], a new training process is automatically initiated. Furthermore, a new model will run for a stabilization period (for example, at least 120 days in this research) before the retraining will be triggered to obtain sufficient statistics to reevaluate the model performance and avoid too frequent a model switch. This continuous learning cycle can renew the model as soon as it is needed, and hence maintain the overall quality and performance of the model. Predefining

[12] This threshold is arbitrary in this context. We do not find a satisfactory approach to determine this objective through data or modelling. Instead, we see this as a decision that may vary among different applications. Given the error rate in the existing literature, we choose this number. When we vary this number, it doesn't affect the poor performing model as this is a relatively high threshold for this application.

the rule of retraining would also make reduce the need for human intervention when market condition changes (less regret or overreactions) and make the model update trackable.

In terms of the number of data points used in each of the steps, for the current application, at the end of 2009, we obtain 4000 data points of which 90% were for the k-fold training and 10% for validation. During the training phase, we use 5-fold cross-validation. We then implement the selected algorithm from 2010 in the out-of-sample closed-loop learning. For reporting the results, instead of presenting only the best model selected at the end of 2009, we report the results of all the best models for different algorithms. This is to examine the performance of our modelling framework.

4 Forecasting Performance

We organize our empirical results into three main sections focusing on prediction accuracy, economic evaluation, and source of predictability. In this section, we study the training, validation and forecasting performance measured by the forecasting accuracy and market timing. In Section 5, we study the economic significance of the forecast. In Section 6 we study the source of predictability through several experiments.

4.1 Training and validation accuracy for modelling at the end of 2009

We start by examining the prediction accuracy of the selected models for all the algorithms at the end of 2009. These are the outputs taken from the "Step 2 validation and algorithm selection". Specifically, after the K-fold cross-validation training, we keep the best models and apply these models to the 400 validation points. Figure 2 reports the training and validation accuracy for each algo. We have four notable results. First, NB has the lowest accuracy suggesting more complex algos add value to this application. Second, a linear model such as LR produces reasonable accuracy. This suggests that an important part of the predictability is coming from the economic relevance of the features we selected in our analysis.

Third, there is a trade-off between the complexity and stability/variability of the model when comparing the in-sample training and the out-sample validation performance. A large drop in the out-sample performance is a suggestion of overfitting. In this regard, the neural network-type model, such as the MLP, has the most complex nonlinear structure. The results of MLP indicate an alarming sign of overfitting with an in-sample accuracy as high as 94%. Nevertheless, its out-of-sample validation accuracy rate of 62.2% is still very good compared to that of NB. Finally, AB provides the best validation results. Interestingly, its validation results are higher than the in-sample results. Should we make the algo choice of decision following our framework at the end of 2009, AB would be the one we pick for implementation in Step 3.

<Insert Figure 2>

4.2 Out-of-sample implementation accuracy between 2010 and 2020.

We report the box plot of the yearly correct ratio for the out-of-sample forecast by algorithms in Figure 3. The mean of the ratio is also reported in Table 2. Consistent with the validation results, Figure 3 shows that the Naïve Bayes (NB) has the lowest accuracy rate while the Decision tree (DT), Random Forest (RF) and AdaBoost (AB) have relatively higher accuracy rates, which are consistently higher than 50%. The MPL performs poorly. This further confirms the potential impact of overfitting identified in the validation stage. The ensemble (ENS) has a performance that is in between DT and RF. It seems to be able to reduce the variability of the model performance among different years (a narrower interquartile range). Overall, this finding confirms that simple probabilistic classifiers (NB) and the most complex method (MPL) produce poor directional forecast, and the decision tree type models seem to be the best tool for this type of classification task.

<Insert Figure 3 >

Finally, we report the statistics of the model *retraining* during the closed-loop learning in Step 3 assuming that we pick each of the best models from each algo and start the closed-loop implementation. By design, the weak algos (in terms of performance) will be retrained more often than the stronger ones. In Figure 4 we see that NB has the highest number of retraining, 23 times in these 11 years. Note that we require a model to run a minimum of 120 days (about half-year in terms of trading days). This suggests that the NB is retrained as soon as 120 days is expired as its performance never be as good as the threshold of error rate at 42.5%. In other words, retraining would not solve the performance problem for a weak algorithm. DT and AB have a relatively lower number of retraining being retrained slightly less than once a year. Finally, the benefit of an ensemble seems to be producing a more stable model that requires the least number of retraining (7 times in 11 years).

Examining the variation of the model performance, even though NB requires a lot of retraining, its performance has low variations between models, but they are invariantly low. By contrast, RF and MLP have a huge variation in their training performance among different stages/models. The training and validation picture for all models is consistent with what we found at the end of 2009 (Figure 2). This is good to see as it suggests that our training regime produces consistent training outcomes for different data sets. Especially the overfitting problem of MLP persists in the close-loop training[13].

<Insert Figure 4>

4.3 Statistical test

Table 2 reports the accuracy and timing measures in the out-of-sample implementation phase[14]. To test the difference in accuracy rates between

[13] In our initial study, we have also included the Support Vector Machine (SVM) in our model choice. However, it seems that this algorithm tends to produce one-sided prediction with close to zero timing ability. We present the results and brief discussion in an online appendix.

[14] To measure market timing. We consider the following measure (Bodie, Kane and Marcus 2018, Chapter 24): Market timing ratio = true positive ratio + true negative ratio -1 . When this ratio is equal to 1 it indicates perfect timing. Asymmetric performance for up or down market conditions will result in a much lower market timing. For example, a model that predicts only up will be correct for all realized up predictions while all wrong for down. This will lead to zero market timing. The model performance measured by accuracy or information ratio can be quite good if the market is in a bull year. But such a model's

models, we conduct Diebold and Mariano's (1995, DM) tests. Given the Diebold-Mariano test tends to reject the null hypothesis too often for small samples. Harvey, Leybourne, and Newbold (1997, HLN) propose modified statistics to address this issue. In our main results, we report the HLN statistics. To further benchmark our findings with the existing method in the literature, we report a simple linear forecasting model known as HAR which has been found to perform well in volatility forecasting as discussed briefly in Section 2. Specifically, we conduct the rolling daily HAR model forecast with three variables: the lagged one-day, weekly average, and monthly average values of VIX. Similarly, to our main analysis, we take 4000 observations as the rolling window.

Table 2 has the following findings. First, Panel A confirms that all but the NB and MLP models outperform the HAR model. Since both HAR and LR can be considered as a 'linear' forecasting model, the key difference between these two approaches is mainly in the number of features included. The outperformance of LR, compared to HAR, further confirms that the additional economic variables included in this study are important to increase the forecasting performance. Among the models, DT and AB are better than the rest in terms of the accuracy rate which is also confirmed by the information coefficients. For the market timing ratio DT, AB, RF and ENS all had above 10% market timing with robust statistical significance. By contrast, the NB, MLP and HAR produce low market timing[15].

<Insert Table 2>

Second, Panel B report the pairwise accuracy tests. It shows that DT has the best accuracy rate compared to all other models. The ensemble model can only beat the NB and MLP models.

performance will have high variations among different periods as the market conditions change.

[15] Note that the information ratio and market timing ratio produce similar conclusion in terms of cross model comparison. This is partly because these training has been implemented with a 'balanced' sample. We demonstrate the usefulness of the market timing ratio in Section 6.4.

Overall, the decision tree family models: DT, RF and AB have the best prediction accuracy. These models consistently beat the HAR model for predicting the directional movement of the VIX.

5 Economic evaluation: a simulated strategy

Diebold and Mariano (1995) pointed out the importance of recognizing the fact that the economic loss associated with a forecast may be poorly assessed by the usual statistical metrics. "… forecasts are used to guide decisions, and the loss associated with a forecast error of a particular sign and size is induced directly by the nature of the decision problem at hand." In the context of our research, the best approach to understanding the economic gain and loss of the forecast is to apply the directional forecast to a trading strategy. This will quantify the size of the consequence of the forecast error in economic terms directly.

To demonstrate the economic significance of the directional forecast, we consider taking the prediction as a trading signal for a long-short strategy. The daily signal can be generated at the market close (3:15 US central time). We trade the signal at the closing VIX price and hold till the next rebalance. We rebalance whenever the signal changes its direction. We are aware that VIX is not directly tradable. For this section, we consider this simulated strategy as a size-weighted signal accuracy test[16].

We study the out-of-sample 'return' of such a strategy. Figure 5 reports the distribution of the mean daily return for different algorithms in the implementation phase. The variations are based on the difference in the statistics over the 11 years. In general, a more accurate prediction produces a higher return. However, there are exceptions among those having relatively higher accuracy, the DT, RF and AB models. Although DT has the highest prediction accuracy it has a relatively lower return compared to RF and AB. And these differences are statistically significant. For example, when we test the difference between DT and AB the t-value is 3.91.

[16] We demonstrate two more realistic investment tests with some tradable instruments in an online appendix.

<Insert Figure 5>

When comparing the risk of the strategies in Table 3, ENS, RF and AB produce high Sharpe ratios which suggest a relatively consistent performance year on year. The AB produces the lowest drawdown.

<Insert Table 3>

Examining the size of the return, most of the models produce an average daily return that is above 50 basis points. If one annualized this by 250 days, it is equivalent to a 125% return annually. Such returns are economically significant but not directly obtainable for several reasons. Two of them are very important to note. First, VIX cannot be directly traded and can only be traded through its futures contracts or other derivatives which are mainly constructed based on the VIX contracts instead of the VIX spot. This creates further 'tracking' errors and derivative risk that would affect the model performance when applying to tradable instruments. Second, it is the transaction costs, especially given the relatively high frequency of rebalancing.

Although the VIX prediction cannot be directly traded, the 'economic' relevance analysis in this section has two important implications. First, the 'return' of this strategy can be considered as a 'weighted' accuracy rate with the size of the VIX movement as the weight for each signal. A large annualized return suggests that the model gets it right more often when the market movement is bigger and therefore economically more important. Second, being able to predict VIX with a meaningful size of movement can be used for economic decisions other than direct trading. For example, this prediction can be used as a further signal in optimizing derivative portfolios; it can also be economically relevant to market makers of SPX and VIX futures and options who can use this directional prediction to improve their market-making in terms of mid-price and spread setting[17]. In the last section of this paper, we are going to explore direct applications of the prediction to various VIX derivatives taking into consideration transaction costs.

[17] We leave the precise form of application to the future research. Any suggestions are most welcome.

6 Source of predictability

In this section, we study the source of predictability with several experiments. We study how these models perform and adapt around large volatility episodes, namely the volatility spikes (6.1). We examine the relevance of economic variables through variable importance analysis (Section 6.2). We then explore how two key features of our model setup affect model performances: the closed-loop training (6.3) and the balanced sampling method (6.4). We study if multi-categories predictions would outperform the binary predictions (6.5). We consider the persistence of the prediction by examining the effect of delay in the predictors (6.6).

6.1 Volatility Spikes

Large episodes of volatility in the market are, in general, caused by unexpected events such as the flash crash of 2010 and the surge of activities in the event of GameStop's social trading in 2021. These spikes in VIX create a significant tail risk in shorting VIX. Since these innovations are truly exogenous to the system in that there is less warning or information embedded in the data. If market participants do learn from each spike and adjust their trading behavior accordingly, these behaviors will be reflected in the data point and our model will be able to take such information into account. However, the next episode of the spike may still catch the system by surprise as it is something that not many participants, if any, in the market would have a foresight of it[18].

To understand how the model prediction is related to this potential tail risk, we study the performance of the models around these episodes of volatility spikes. Table 4 reports a summary of the model performance on the days that VIX changed by more than 20% either way including negative and positive spikes.

[18] Therefore, we do not expect that our system would be able to pick up those large jumps in volatility. If it happens to get it right, we consider it lucky. What concern us is the downside risk brought by these volatility jumps. One comprehensive approach is to develop a system that can predict the dump. To this end, we do not find a viable solution as these are truly unknown to the general public and the cause can be different on different occasions. Furthermore, the low number of occurrences makes it ineffective for ML to learn from the realized data even if there are any leading indicators presented in the data.

<Insert Table 4>

We can see there are many more positive (64) than negative spikes (10) during our 11-year out-of-sample period. For negative spikes, the error rate is low for most of the models with a maximum of 20%. By contrast, positive spikes are obvious surprises to all models with error rates often much higher than 50% except for the NB model. This confirms that positive VIX spikes are indeed a concern to the application of these models' prediction[19]. Nevertheless, the average returns suggest that the overall impact of these spikes diversified away over these 11 years.

Although the spikes are surprises to most of the models, what happens after the spike provides further information about the effectiveness of the models in adapting to the new environment through the new data points coming into the model. Table 5 shows the mean initial losses for all those *incorrect* predictions on the spike days for each model. On average the initial daily loss is slightly above 30%. We track the return following these initial spikes and report the cumulative P&L until 20 and 60 days after the spike (including the initial losses)[20]. We calculate the recovery ratio by comparing the gains made during the subsequent period to the initial losses. It shows that the decision-tree-type models did relatively better after these spikes. For example, 20 days after the initial spikes the AB model can recover 93% of the initial losses while the HAR model only recovers 23%. By 60 days, most models were covered fully except for NB and MLP.

<Insert Table 5>

Overall, although VIX spikes are not predictable to the system the decision-tree-type ML algorithms can adapt quickly by abstracting new information from the data to recover from the losses.

[19] Large reduction of volatility (negative spikes) is less exogenous than large increase of volatility (positive spikes) is because most of the negative spikes are as a result of reversal from the positive spikes.

[20] This cumulation is done by summing the daily return without compounding. Such a return can be achieved by investing a fixed amount of capital into the strategy every day. Similar to abnormal return calculation using sum or average of the return in this case is more relevant to evaluate the expected outcome without further complicate by the order of the return in the series (Fama, Fisher, Jensen, and Roll, 1969)

6.2 Variable importance and variable selections

One of the key trends in ML is the importance of interpretable ML. The availability of statistics such as variable importance helps researchers to understand, to some degree, the source of the predictability. This may help uncover new important variables of determinants that have been overlooked by the literature when traditional methods are applied.

In this research, we use a forest of trees to evaluate the importance of features on the VIX signal classification task. An ensemble learning method based on decision trees, ExtraTreeClassifier, is selected to use a meta estimator that fits several randomized extra-trees on various sub-samples of the VIX training dataset. Then the feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node, which builds the indication of how much the prediction depends on the feature.

Table 6 reports the top 20 variables according to their variable importance ranking. Panel A reports the statistics from the training data at the end of 2009 before the out-of-sample implementation. Panel B reports the ranking and variables based on the average variable importance calculated for the nine retraining for the AB model during the implementation stage. In both panels, we also report the cross-referenced ranking from the other panel to examine the consistency of the ranking in different stages of the modelling.

Table 6 shows that the weekly jobless report plays a decisively important role in all models[21]. This is something that hasn't been explicitly featured in the discussion of volatility study. It suggests that the most important variable to predict market fear is potentially the fear of unemployment. This could potentially be due to the heterogeneity in the interpretation of this statistic which leads to high predictability in the next period's volatility.

<Insert Table 6>

[21] Weekly initial unemployment claims track the number of people who have filed jobless claims for the first time weekly with the appropriate government labour office. This number represents a net inflow of people receiving unemployment benefits.

For other variables, seasonality variables such as the day of the week, day of the month or day to the next VIX contract expiry (Wednesday) also have a relatively high contribution. These variables capture the pattern of investor behavior that is driven by seasonality in the real economic cycle. For example, much macroeconomic news is released on Thursdays.

The rest of the variables are mainly technical by nature. For example, the Relative Strength Index (RSI) of SPX and VIX are highly ranked. Another important technical indicator group are the SPX's member statistics such as the percentage of members with new 52-week highs. Nevertheless, the commodity market such as oil (CL and CO1) and gold (AUX) also play some important roles.

Finally, the last columns of the two panels show that except for the top two variables, there are large variations in the variable importance between models. In an untabulated test, we find that the correlations of the variable importance in the later models with the initial model range from 63% to 77%. This suggests some consistency in the importance of variables while indicating there are time variations. This confirms the importance of updating the model training.

Overall, the findings in the top 20 variables confirm the technical nature of these short-term prediction exercises. The underlining source of predictability would follow a similar argument for technical analysis. It is most likely due to the reversal or momentum effect driven by the behavioral or liquidity condition of the market. Nevertheless, economic variables such as the Jobless claim and commodity are also important sources to predict the next day's VIX movement, which is most likely driven by heterogeneity in the interpretation of the new information.

However, it would be misleading to conclude the importance of a type of the variable by looking at just the most influential variable. They only have a relatively higher variable importance. The overall model performance is driven by all the variables. To see this, we summarize the mean, minimum, maximum and sum of the variable importance by category. To be concise in Table 7, we report 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.

<Insert Table 7>

The sum of variable importance confirms the source of predictability is from both technical groups (the SPX and VIX Techs) and some fundamental such as macroeconomic variables. The mean of the variable importance in Table 7 shows the average relevance of the variables included in each category. It shows that variables in the seasonality category all have relatively high variable importance. The next two groups are commodity and currency variables. The sum of the variable importance suggests that even though individually the contribution is low in those categories, such as Macroeconomics, collectively they provide important information for the models. Importantly, the VIX Techs group contribute only 11.46% out of the 100% among the variables. This suggests that the existing studies using only VIX's historical data such as the HAR model will have missed a large number of potential explanatory variables.

6.3 One-time model vs dynamic continuous learning model

To demonstrate the advantage of the continuous learning framework, we experimented with a one-time model approach. The model is built only once without further updates during the implementation stage. The same model is applied to the whole out-of-sample period (11 years). A comparison of the performance is given in Table 8.

We can see that without the dynamic retraining, among the onetime models, only those more advanced models such as RF, AB and MLP can provide average forecast accuracy that is greater than 50%. Consequently, the dynamic retraining setup provides less improvement for these models. The HLN tests so that the improvement in the AB and MLP is not statistically significant. The largest improvement brought by dynamic learning is in the DT model which has a massive improvement from 49% to 58%. It also boosts the Ensemble

significantly. For the NB models, a more frequent training setup makes its performance worse.

Overall, these findings suggest updating the learning model with new information on demand provides a significant improvement over the static model, especially for the decision tree models.

<Insert Table 8>

6.4 Balanced vs Unbalanced sampling

One of the concerns when using a nonlinear model for a classification task is that the model would produce a one-sided prediction as the prediction accuracy may turn out to be good in the training sample. However, such a model can be very wrong when it is applied out of the sample. The timing ratio measurement was designed to measure the model's balance in predicting both up and down signals.

The main results we have presented so far have included the important data engineering to construct a "balanced" sample. We select 4000 data points from previous years keeping an equal amount of up and down observations. This means we will use more than 4000 data to construct the sample given the potential of unbalanced outcomes in the raw data. In this section, we compare the results of the unbalanced and balanced approach for the AB model.

Table 9 reports the model statistics. Comparing these with the main results, the unbalanced sample has a higher correction ratio and slightly higher information ratio. However, taking into consideration the accuracy when it is up or down, the timing ratio in the unbalanced model is much lower than the main results. Furthermore, the economic importance of balanced predictions is demonstrated in Panel B. We see that the before-cost simulated return is much lower in the unbalanced model (49 basis points) than in the balanced model (90 basis points). An unbalanced sample training also comes with a higher drawdown as well.

Overall, our findings in this session demonstrate the importance and benefit of providing the model with a balanced number of observations for both

up and down realizations so that the machine would have an equal amount of information to differentiate these two outcomes. It reduces the likelihood of an unconditionally one-sided prediction and improves the accuracy of both directions' prediction. Such models are more robust to various market conditions. The improvement is economically significant.

<Insert Table 9>

6.5 The size of VIX changes and multi-category forecasting

One of the limitations of a binary forecast is that it hasn't taken into consideration the potential magnitude of the next period's movement as part of the objective. A high accuracy rate may not lead to an economically meaningful prediction if it often gets wrong when there are large changes in the market and corrects only when there are small changes in the market. Our simulated investment strategy in Section 0 has demonstrated the significance of our prediction through the return-weighted signal quality. In this section, we further explore the relationship between directional predictions and the magnitude of VIX movements in two ways.

First, we develop a measure that is similar to the market timing ratio to examine the 'size timing'. Specifically, we group the realized outcome into big and small sizes of change compared to the historical rolling distribution of the change in the past 250 observations. The threshold of big changes in the upper and lower 15 percentiles. Post estimations, we then examine the accuracy of these two-size groups: big and small. Second, to study the potential benefit of taking into consideration the size of the expected change, we conducted a multi-category prediction experiment. We extend our original binary prediction into a four-category prediction model: up-small, up-big, down-small, and down-big (referred as to 4D).

Table 10 reports the findings for the AB model. In this post-estimation analysis, we calculate the correct ratio for the 4D to be comparable with 2D predictions in that we only count the accuracy for up and downs without considering the size category. It shows that if one uses the 4D signals it doesn't improve the overall directional prediction accuracy. The overall accuracy rate is

similar to those in the 2D models. And surprisingly, the market timing ratio is worsened in the 4D model. Furthermore, the size timing analyses show that the binary model has done quite well in capturing the bigger size movement than the smaller size movement.

Overall, these findings don't find the advantage of conducting a finer category forecast. One potential reason for this could be that when four instead of two categories are needed to learn we should double the sample size so that there are comparable observations for the model to learn from as in the 2D forecast for a one given outcome category. This is not feasible currently given the total size of our sample.

<Insert Table 10>

6.6 Persistence of the prediction

We examine the persistence of the prediction to answer two questions. First, in practice, some of the data may be delayed (e.g., equity data is delayed by up to 15 minutes for a standard subscription to Bloomberg) or required a higher cost to obtain real-time feeds. A study of the signal relevance with some time lag can help with the decision on the choice of data feed requirement for implementing the models. Second, testing signal persistence would also provide information about the stability of the model and further reveal the source of predictability from the model. For example, a gradual decline in predicting accuracy would provide some confidence that the model is indeed using the most recent data to predict the immediate future efficiently and such power will be diminished if there is a longer gap between the input data and the objective signal.

Specifically, we carry out three tests. First, we study the accuracy of the signal applied with one day delay. This is a 'big' gap given our objective is daily prediction. Second, we study the accuracy of the signal on the next day's open-to-open VIX movement. Third, instead of using the signal generated by the model using the close-to-close VIX price change in the second experiment, we directly model the prediction for the next day's open-to-open VIX price change using data from the current day's close. This allows us to do the modelling

before the VIX market opens at 2 am US central time (the beginning of the global trading hour).

Table 11 shows the results of our different experiments with the AB model. The first row is the benchmark results as reported in the main result. Applying the signal with one day delay would reduce the mean forecast accuracy by nearly 3%. The information and timing ratio also reduced significantly. Similarly, applying the close-to-close prediction to the next day's open-to-open return would reduce the accuracy. For these two experiments, Panel B also shows that the reductions of the return are significant with about 70 basis point drops. These findings suggest that our model produces a timely forecast that reflects an immediate change in market condition and the timeliness of using the signal is very important.

If one considers delaying the model prediction due to data availability or cost, a possible way to achieve this is to predict the next day's open-to-open price movement. When we model the next day's open-to-open directly and apply this signal to the open-to-open return, the model performance is much better than the other two experiments although it is still slightly worse than the original benchmark model in terms of the model accuracy. The return performance, although lower than the benchmark, still generates 69 basis points daily. Overall, this last experiment confirms the features selected and the model structure are robust to a time gap between the market close and the next open.

Furthermore, the variation of the model performance is expected in that timelier information would produce a better forecast and when the training and application targets are in line the performance is better than when they are mismatched. All of these confirm that the features and the model structures are efficient in the use of the information it has to predict the target specified.

<Insert Table 11>

7 Conclusions

We study if VIX is predictable and what is the source of predictability. We show that daily VIX can be predicted at an accuracy that is better than existing evidence in the literature and economically significant. We demonstrate the source of predictability coming from two main areas. First, it is still the most important role to play by humans which is the choice of economically relevant variables and the translation of the prediction task into a form that the machine can be interpreted. To this end we include large coverage of economic and financial data in our study and especially uncover the most important variable that is used by the machine to predict VIX is the weekly jobless claim data. This new finding contributes to the existing literature on the determinant of volatility and market fear. The linkage between the labor market and overall market volatility hasn't been covered sufficiently in the current literature. Further research may look into the potential channel of this impact through theoretical and empirical studies.

Second, the source of the predictability comes from the forecasting framework. We developed a three-step automated and adaptive training framework based on AutoML with explainability and trackability. This framework has several districted features that aim at tackling some of the key challenges in applying ML to financial forecasting. 1) The framework utilized the AutoML HPO with Grid-Search and K-fold cross-validation to reduce human intervention during the training. It enables automation and produces a trackable outcome. The replacement of the conventional manual hyperparameter tuning hyperparameter for each model manually by an automated Grid-Search with the K-Fold validation method over the pre-defined parameter range improves the efficiency and consistency of the algorithm selection and model tuning process. 2) the out-of-sample validation after training ensures the robustness of the trained model and avoids overfitting or look-back bias. This avoids false dictation due to multiple testing on the same set of data (Harvey and Liu, 2015). 3) the proactive monitoring of performance drift and model decay and the close-loop retraining process make the model switch decision more systematic and automatic. This reduces human error and trackable model switching for audit purposes.

Through the process of this research, one lesson that we learned from the machine is that it cannot learn from what is not in the data. Especially, one needs to recognize the limitation of the method. For example, the true spectacular movement in the VIX would not be predictable (at least with public information we gather from the financial market), what the system can do is deduce some under or overreaction to the 'old' news. Our results show that there is a robust and consistent pattern of behavior that the machine can learn, and the predictions are practically relevant.

References

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Paul Labys, 2003, Modeling and forecasting realized volatility, *Econometrica* 71, 579–625.

Bali, T. G., and H. Zhou, 2016, Risk, uncertainty, and expected returns. Journal of Financial and Quantitative Analysis 51, 707–735.

Ballings, Michel, Dirk van den Poel, Nathalie Hespeels, and Ruben Gryp, 2015, Evaluating multiple classifiers for stock price direction prediction, *Expert Systems with Applications* 42, 7046–7056.

Bodie, Zvi, Alex Kane, and Alan J. Marcus, 2018, *Investment* (McGraw Hill).

Bollerslev, Tim, 1986, Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* 31, 307–327.

Booth, Ash, Enrico Gerding, and Frank McGroarty, 2014, Automated trading with performance weighted random forests and seasonality, *Expert Systems with Applications* 41, 3651–3661.

Booth, Ash, Enrico Gerding, and Frank McGroarty, 2015, Performance-weighted ensembles of random forests for predicting price impact, *Quantitative Finance* 15, 1823–1835.

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

Bucci, Andrea, 2020, Realized Volatility Forecasting with Neural Networks, *Journal of Financial Econometrics* 18, 502–531.

Corsi, Fulvio, 2009, A simple approximate long-memory model of realized volatility, *Journal of Financial Econometrics* 7, 174–196.

David, Alexander, and Pietro Veronesi, 2002, Option Prices with Uncertain Fundamentals: Theory and Evidence on the Dynamics of Implied Volatilities, *Working paper, University of Chicago*.

Degiannakis, Stavros, George Filis, and Hossein Hassani, 2018, Forecasting global stock market implied volatility indices, *Journal of Empirical Finance* 46, 111–129.

Diebold, F. X. and R. S. Mariano, 1995, Comparing predictive accuracy, *Journal of Business and Economic Statistics*, 13, 253-63.

Donaldson, R. Glen, and Mark Kamstra, 1997, An artificial neural network-GARCH model for international stock return volatility, *Journal of Empirical Finance* 4, 17–46.

Engle, Robert F, 1982, Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation, *Econometrica* 50, 987–1007.

Fernandes, M., M.C. Medeiros, and M. Scharth, 2014, Modeling and predicting the CBOE market volatility index, *Journal of Banking and Finance* 40, 1–10.

Guidolin, Massimo, and Allan Timmermann, 2003, Option prices under Bayesian learning: Implied volatility dynamics and predictive densities, *Journal of Economic Dynamics and Control* 27, 717–769.

Hamid, Shaikh A., and Zahid Iqbal, 2004, Using neural networks for forecasting volatility of S&P 500 Index futures prices, *Journal of Business Research* 57, 1116–1125.

Harvey, Campbell R., and Yan Liu, 2015, Backtesting, *Journal of Portfolio Management* 42, 13–28.

Harvey, D., S. Leybourne, and P. Newbold, 1997,. Testing the equality of prediction mean squared errors, *International Journal of Forecasting* 13, 281-91.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, *An Introduction to Statistical Learning: With Applications in R* (Springer Science & Business Media, New York).

Jurado, K., Ludvigson, S. C., and Ng, S., 2015, Measuring uncertainty. *American Economic Review* 105, 1177–1216.

Konstantinidi, Eirini, George Skiadopoulos, and Emilia Tzagkaraki, 2008, Can the evolution of implied volatility be forecasted? Evidence from European and US implied volatility indices, *Journal of Banking and Finance* 32, 2401–2411.

Kristjanpoller, Werner, and Marcel C. Minutolo, 2018, A hybrid volatility forecasting framework integrating GARCH, artificial neural network, technical analysis and principal components analysis, *Expert Systems with Applications* 109, 1–11.

Maciel, Leandro, Fernando Gomide, and Rosangela Ballini, 2016, Evolving Fuzzy-GARCH Approach for Financial Volatility Modeling and Forecasting, *Computational Economics* 48, 379–398.

Paye, Bradley S., 2012, "Déjà vol": Predictive regressions for aggregate stock market volatility using macroeconomic variables, *Journal of Financial Economics* 106, 527–546.

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

Rasekhschaffe, Keywan Christian, and Robert C. Jones, 2019, Machine Learning for Stock Selection, *Financial Analysts Journal* 75, 70–88.

Rhoads, Russell, 2011, Trading VIX Derivatives: Trading and Hedging Strategies Using VIX Futures, Options, and Exchange-Traded Notes, *John Wiley & Sons, Inc.*

Rhoads, Russell, 2020, The VIX Trader's Handbook: The history, patterns, and strategies every volatility trader needs to know, *Harriman House*.

Whaley, Robert E., 2000, The investor fear gauge: Explication of the CBOE VIX, *Journal of Portfolio Management* 26, 12–17.

Figures and Tables

Figure 1. The adaptive continuous learning methodology

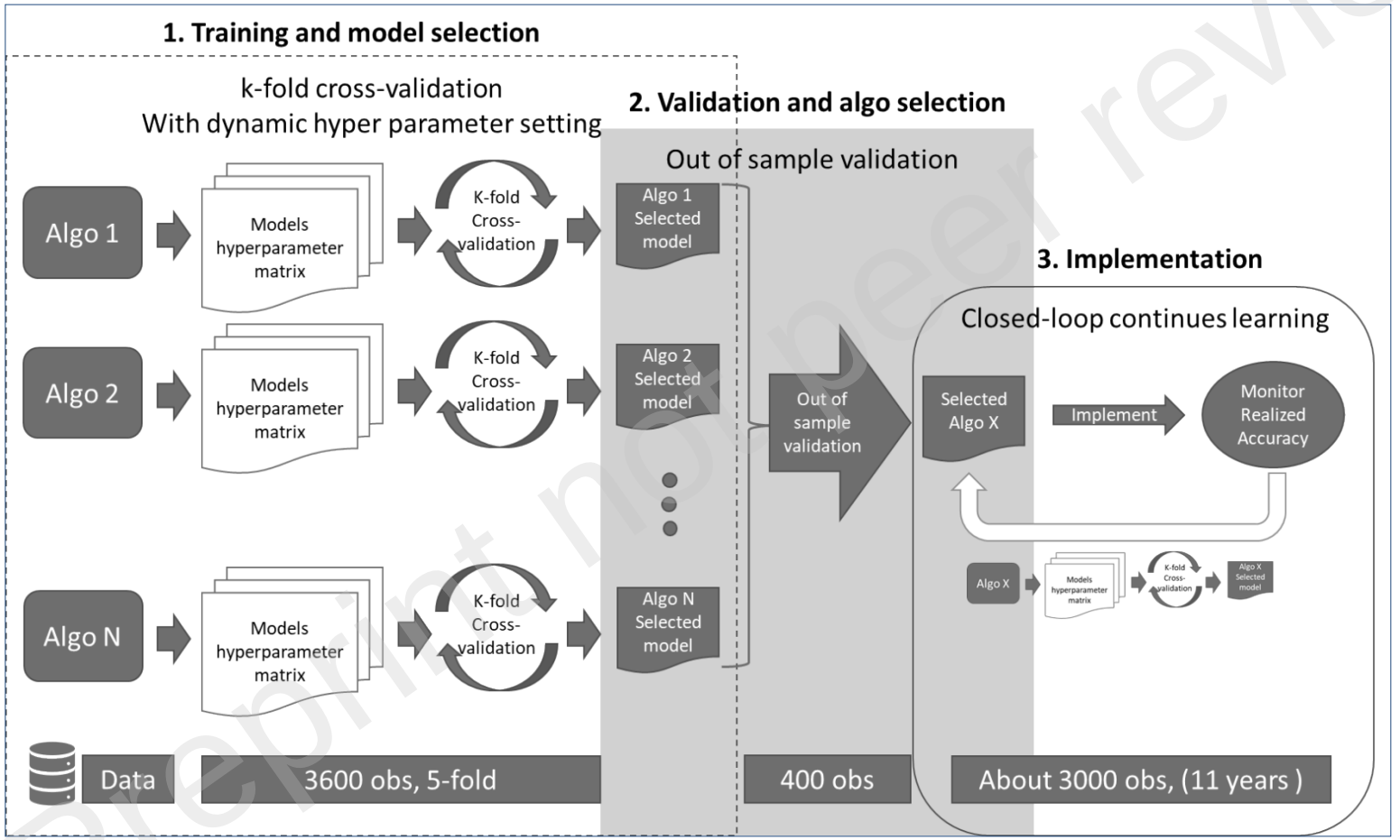


Figure 2. Training and validation accuracy for modelling at the end of 2009

This figure reports the accuracy ratios for the training and validation of each model.

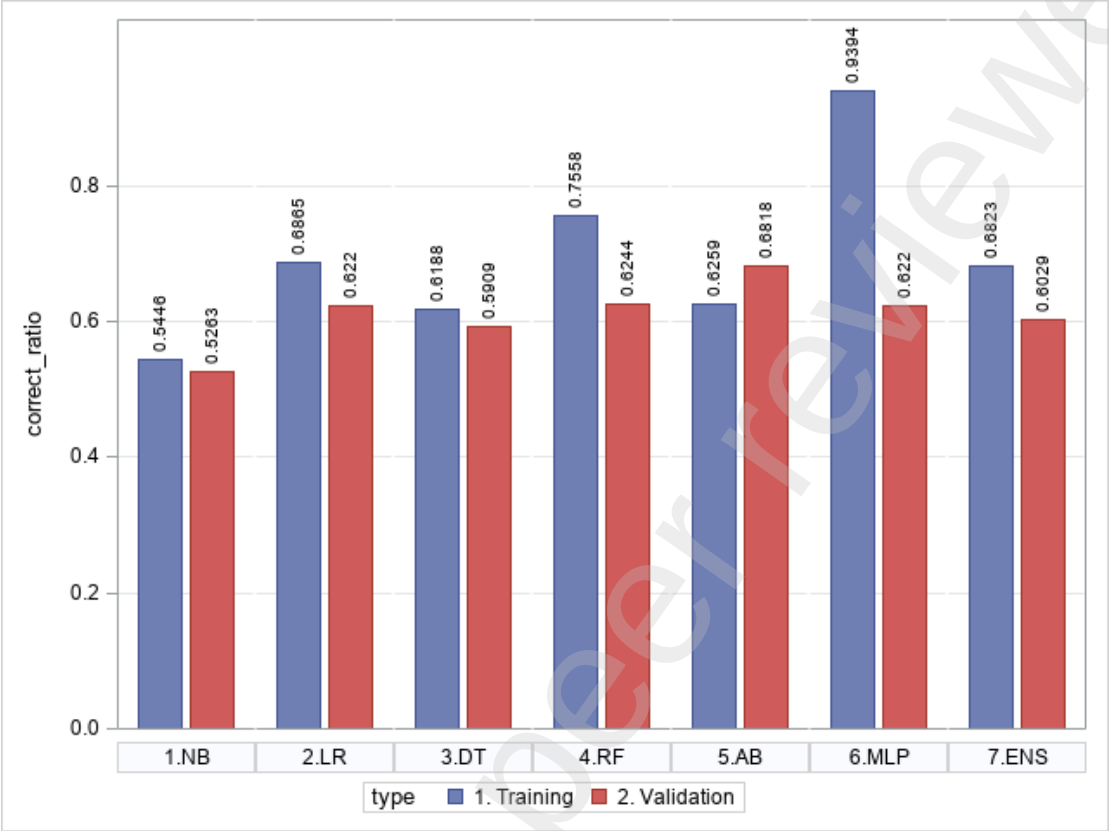


Figure 3. Out-of-sample implementation accuracy by model

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.

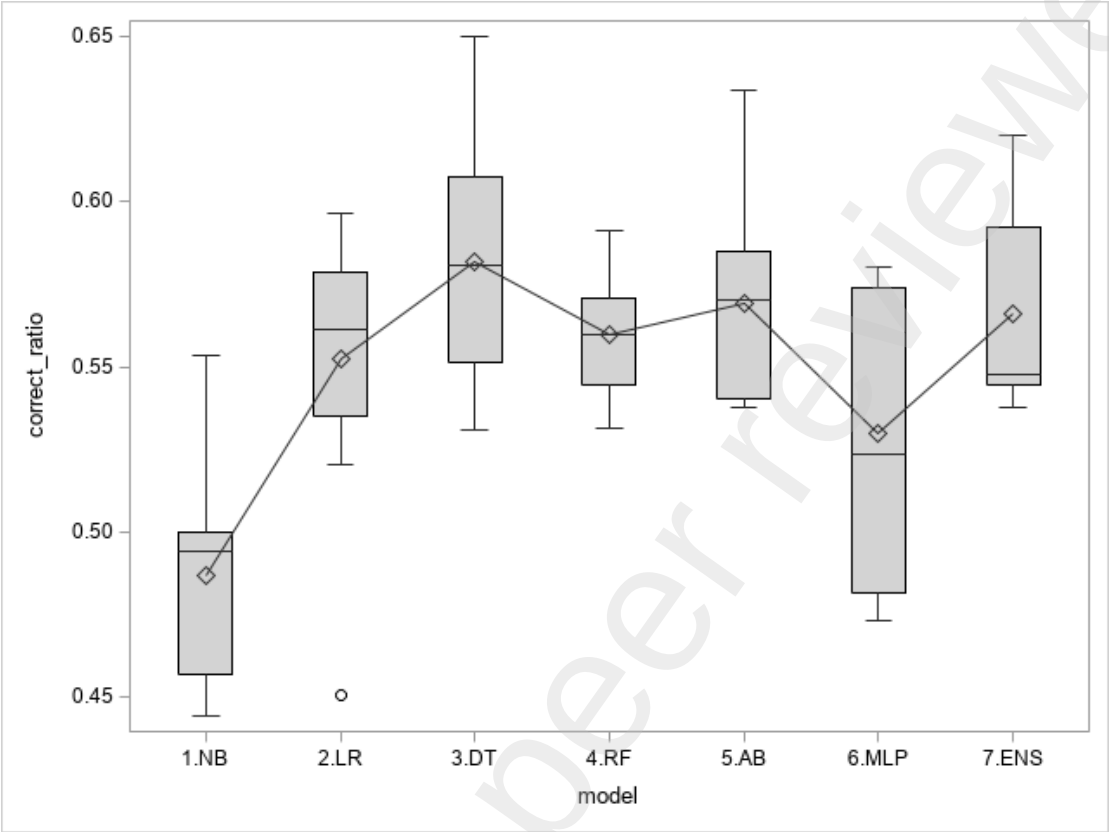
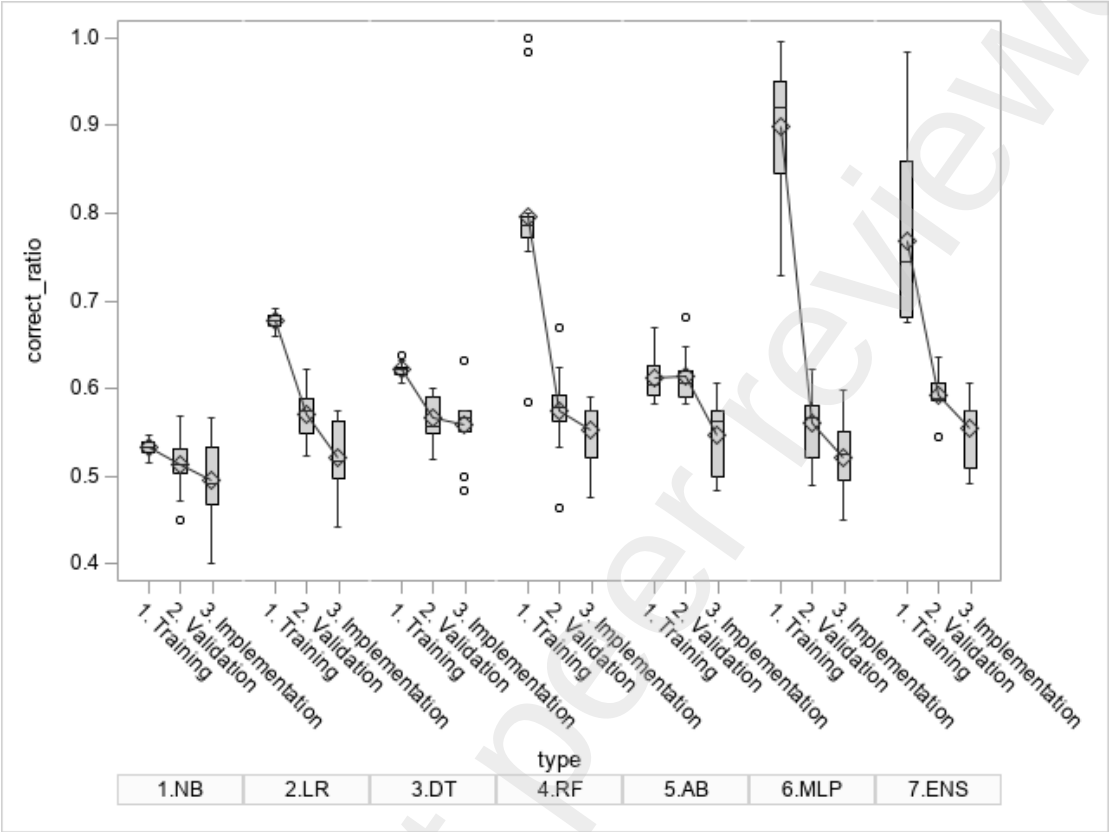


Figure 4. Retraining, validation, and implementation accuracy during the close-loop implementation period

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.



	1.NB	2.LR	3.DT	4.RF	5.AB	6.MLP	7.ENS
number of retraining	23	20	9	16	10	16	7

Figure 5. 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.

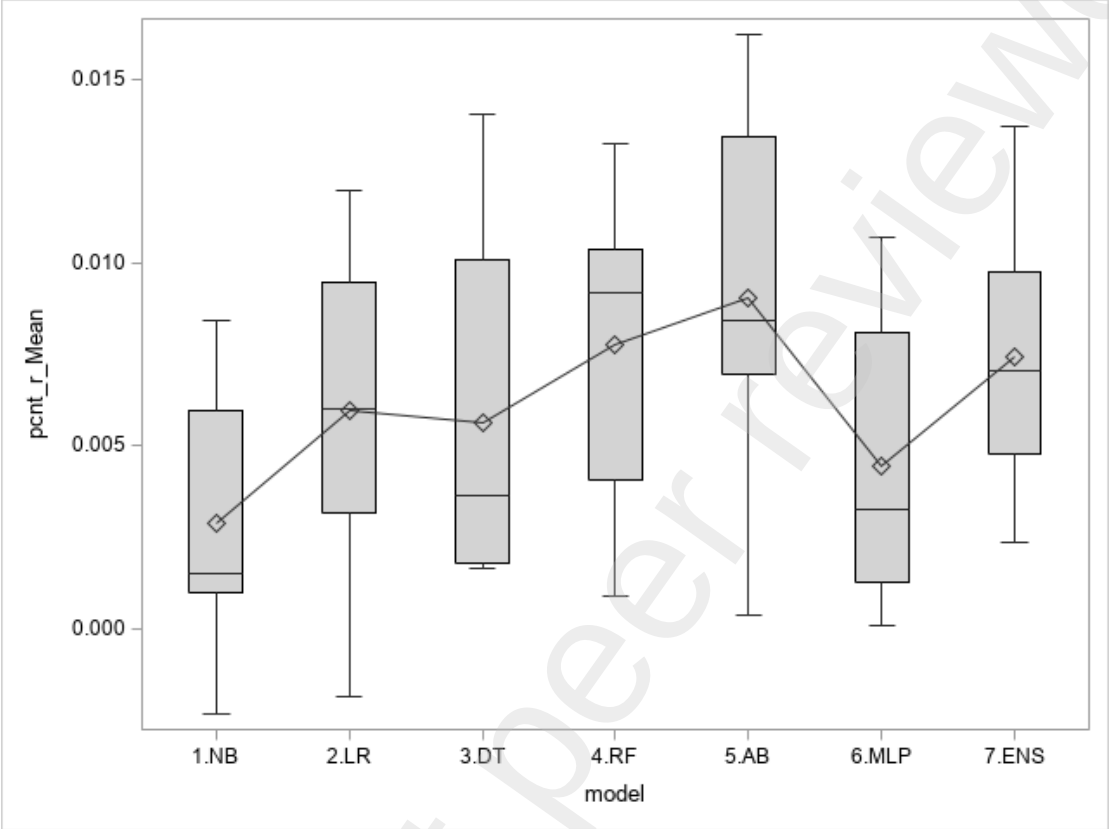


Table 1. Summary of variables by groups

Groups	Examples	Number of fields
SPX Member Tech		26
	PCT MEMB PX GT 50W MOV AVG	
	PCT MEMB WITH 14D RSI LT 30	
	SPX Index PCT MEMBERS WITH NEW	
SPX Options and Futures		14
	Total Put Volume	
	Total Call Open Interest	
	Aggregate Volume of Futures Contracts	
SPX Subindex		33
	S5BANKX Index	
	S5RETL Index	
	S5AUCO Index	
SPX Tech		45
	Average Volume 5 Day	
	ARMS Daily Index	
	Volatility 200 Day	
Vix Tech		34
	RSI 3 Day	
	Moving Average 5 Day	
	Max30	
World Equity Index		18
	DAX Index	
	CCMP Index	
	UKX Index	
Major Equities		12
	IBM US Equity	
	AAPL US Equity	
	GE US Equity	
Macroeconomic		61
	PPI CHNG Index	
	RSTAMOM Index	
	IMP1CHNG Index	
Govt & Corp Bond		14
	CSI BARC Index	
	USGG2YR Index	
	LF98TRUU Index	
Currency		7
	EUR Curncy	
	JPY Curncy	
	GBP Curncy	
Commodity		9
	BCOM Index	
	CL1 COMB Comdty	
	CO1 COMB Comdty	
Seasonality		5
	Day of the week	
	Week of the year	
	Days to next maturity Wednesday	
Total		278

Table 2. Forecast accuracy summary

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, Leybourne, and Newbold (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 test 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.

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***

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

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.

	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***

Table 4. Prediction performance on VIX spikes days

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.

					Return		
Models	**Spikes**	**Number of days**	**Error days**	**% Err**	**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

Table 5. Return performance following the VIX spikes days

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.

Models	Initial losses on spike day (1)	20 days after initial		60 days after initial		N
		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

Table 6. Top 20 Variable Importance by ranking

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

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

Table 7. Variable importance by category for All model summary

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.

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

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

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, Leybourne, and Newbold (1997) test on the difference in accuracy rate between the two training approachs. ***, **, and * indicate statistical significance at 1%, 5% and 10% level respectively.

	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***

Table 9. Results for the unbalanced sample of the Adaptive Boosting model

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.

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

Table 10. Results of size-timing and multi-category prediction of the Adaptive Boosting model.

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.

	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

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

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.

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%

Online Appendix

Online Appendix I. List of variables

ID	Category	Security name	Filed
1	Commodity	Bloomberg Commodity Index	Price Change 1 Day Percent
2	Commodity	WTI CRUDE FUTURE	Price Change 1 Day Percent
3	Commodity	BRENT CRUDE FUTR	Price Change 1 Day Percent
4	Commodity	COPPER FUTURE	Price Change 1 Day Percent
5	Commodity	GOLD 100 OZ FUTR	Price Change 1 Day Percent
6	Commodity	SOYBEAN FUTURE	Price Change 1 Day Percent
7	Commodity	CORN FUTURE	Price Change 1 Day Percent
8	Commodity	SUGAR	Price Change 1 Day Percent
9	Commodity	Gold Spot $/Oz	Price Change 1 Day Percent
10	Currency	Euro Spot	Price Change 1 Day Percent
11	Currency	Japanese Yen Spot	Price Change 1 Day Percent
12	Currency	British Pound Spot	Price Change 1 Day Percent
13	Currency	Australian Dollar Spot	Price Change 1 Day Percent
14	Currency	China Renminbi Spot	Price Change 1 Day Percent
15	Currency	Brazilian Real Spot	Price Change 1 Day Percent
16	Currency	DOLLAR INDEX SPOT	Price Change 1 Day Percent
17	Govt & Corp Bond	US Generic Govt 2 Yr	Price Change 1 Day Percent
18	Govt & Corp Bond	US Generic Govt 5 Yr	Price Change 1 Day Percent
19	Govt & Corp Bond	US Generic Govt 3 Yr	Price Change 1 Day Percent
20	Govt & Corp Bond	US Generic Govt 12 Mth	Price Change 1 Day Percent
21	Govt & Corp Bond	US Generic Govt 3 Mth	Price Change 1 Day Percent
22	Govt & Corp Bond	US Generic Govt 6 Mth	Price Change 1 Day Percent
23	Govt & Corp Bond	US Generic Govt 30 Yr	Price Change 1 Day Percent
24	Govt & Corp Bond	US Generic Govt TII 10 Yr	Price Change 1 Day Percent
25	Govt & Corp Bond	US Generic Govt TII 5 Yr	Price Change 1 Day Percent
26	Govt & Corp Bond	US Corporate High Yield	Price Change 1 Day Percent
27	Govt & Corp Bond	Corporate	Price Change 1 Day Percent
28	Govt & Corp Bond	Corporate	Price Change 1 Day Percent
29	Govt & Corp Bond	UST 13-Week Bill High Discount	Price Change 1 Day Percent
30	Govt & Corp Bond	Ted Spread	Price Change 1 Day Percent
31	Macroeconomic	ISM Manufacturing PMI SA	ISM PMI
32	Macroeconomic	ISM Services PMI	Services PMI
33	Macroeconomic	U-3 US Unemployment Rate Total	Total SA
34	Macroeconomic	US PPI Finished Goods SA MoM%	Goods MoM SA
35	Macroeconomic	Adjusted Retail & Food Service	Monthly % Change
36	Macroeconomic	US Import Price Index by End U	% Change
37	Macroeconomic	US Export Price by End Use All	% Change
38	Macroeconomic	US Real Average Weekly Earning	CES0500000012
39	Macroeconomic	GDP US Chained 2012 Dollars Qo	QoQ % Change Annualized
40	Macroeconomic	US Labor Productivity Output P	PRS85006092
41	Macroeconomic	US Unit Labor Costs Nonfarm Bu	PRS85006112

ID	Category	Security name	Filed
42	Macroeconomic	US Employees on Nonfarm Payrol	Net Change SA
43	Macroeconomic	US Employees on Nonfarm Payrol	Private Chng SA
44	Macroeconomic	US Employees on Nonfarm Payrol	Net Change
45	Macroeconomic	Federal Funds Target Rate - Up	Fed Funds Target Rate US
46	Macroeconomic	US Initial Jobless Claims SA	Initial Jobless Claims SA
47	Macroeconomic	US CPI Urban Consumers MoM SA	MoM % SA
48	Macroeconomic	Conference Board Consumer Conf	Confidence
49	Macroeconomic	US Durable Goods New Orders In	Month % change
50	Macroeconomic	MBA US Mortgage Market Inde	WoW% Change
51	Macroeconomic	US New One Family Houses Sold	Total sold
52	Macroeconomic	US New Privately Owned Housing	US Building Housing Starts
53	Macroeconomic	US Industrial Production MOM S	Month % change
54	Macroeconomic	US Manufacturers New Orders To	Monthly % Change
55	Macroeconomic	US Personal Income MoM SA	MoM % Change
56	Macroeconomic	US Personal Consumption Expend	Monthly % Change
57	Macroeconomic	US Trade Balance of Goods and	US Trade Balance
58	Macroeconomic	Conference Board US Leading In	Monthly % Change
59	Macroeconomic	University of Michigan Consume	Univ. of Michigan Sentiment
60	Macroeconomic	ISM Manufacturing PMI SA	Change
61	Macroeconomic	ISM Services PMI	Change
62	Macroeconomic	U-3 US Unemployment Rate Total	Change
63	Macroeconomic	US PPI Finished Goods SA MoM%	Change
64	Macroeconomic	Adjusted Retail & Food Service	Change
65	Macroeconomic	US Import Price Index by End U	Change
66	Macroeconomic	US Export Price by End Use All	Change
67	Macroeconomic	US Real Average Weekly Earning	Change
68	Macroeconomic	GDP US Chained 2012 Dollars Qo	Change
69	Macroeconomic	US Labor Productivity Output P	Change
70	Macroeconomic	US Unit Labor Costs Nonfarm Bu	Change
71	Macroeconomic	US Employees on Nonfarm Payrol	Change
72	Macroeconomic	US Employees on Nonfarm Payrol	Change
73	Macroeconomic	US Employees on Nonfarm Payrol	Change
74	Macroeconomic	Federal Funds Target Rate - Up	Change
75	Macroeconomic	US Initial Jobless Claims SA	Change
76	Macroeconomic	US CPI Urban Consumers MoM SA	Change
77	Macroeconomic	Conference Board Consumer Conf	Change
78	Macroeconomic	US Durable Goods New Orders In	Change
79	Macroeconomic	MBA US Mortgage Market Inde	Change
80	Macroeconomic	US New One Family Houses Sold	Change
81	Macroeconomic	US New Privately Owned Housing	Change
82	Macroeconomic	US Industrial Production MOM S	Change
83	Macroeconomic	US Manufacturers New Orders To	Change
84	Macroeconomic	US Personal Income MoM SA	Change
85	Macroeconomic	US Personal Consumption Expend	Change
86	Macroeconomic	US Trade Balance of Goods and	Change
87	Macroeconomic	Conference Board US Leading In	Change

ID	Category	Security name	Filed
88	Macroeconomic	University of Michigan Consume	Change
89	Macroeconomic	ICE LIBOR USD 1 Month	Last Price
90	Macroeconomic	ICE LIBOR USD 1 Month	Price Change 1 Day Percent
91	Macroeconomic	US Generic Govt 10 Yr	Price Change 1 Day Percent
92	Major Equities	INTL BUSINESS MACHINES CORP	Price Change 1 Day Percent
93	Major Equities	APPLE INC	Price Change 1 Day Percent
94	Major Equities	AMAZON.COM INC	Price Change 1 Day Percent
95	Major Equities	GENERAL ELECTRIC CO	Price Change 1 Day Percent
96	Major Equities	CELGENE CORP	Price Change 1 Day Percent
97	Major Equities	MICRON TECHNOLOGY INC	Price Change 1 Day Percent
98	Major Equities	MICROSOFT CORP	Price Change 1 Day Percent
99	Major Equities	BRISTOL-MYERS SQUIBB CO	Price Change 1 Day Percent
100	Major Equities	FEDEX CORP	Price Change 1 Day Percent
101	Major Equities	GOLDMAN SACHS GROUP INC	Price Change 1 Day Percent
102	Major Equities	PROLOGIS INC	Price Change 1 Day Percent
103	Major Equities	NVIDIA CORP	Price Change 1 Day Percent
104	Seasonality		Day of the month
105	Seasonality		Day of the week
106	Seasonality		Week of the year
107	Seasonality		Month of the year
108	Seasonality		Day to next expired Wed
109	SPX Member Tech	S&P 500 INDEX	Pct of Members w/Px Below Lower Bollinger Band
110	SPX Member Tech	S&P 500 INDEX	Pct of Members w/Px Above Upper Bollinger Band
111	SPX Member Tech	S&P 500 INDEX	Percentage of Members with Px > 10 Day Moving Avg
112	SPX Member Tech	S&P 500 INDEX	Percentage of Members with Px > 20 Day Moving Avg
113	SPX Member Tech	S&P 500 INDEX	Percentage of Members with MACD > Base Line Zero
114	SPX Member Tech	S&P 500 INDEX	Percentage of Members with Px > 150 Day Moving Avg
115	SPX Member Tech	S&P 500 INDEX	Percentage of Members with Signal > Base Line Zero
116	SPX Member Tech	S&P 500 INDEX	Percentage of Members with Px > 250 Day Moving Avg
117	SPX Member Tech	S&P 500 INDEX	Percentage of Members with Px > 10 Wk Moving Avg
118	SPX Member Tech	S&P 500 INDEX	Percentage of Members with Px > 50 Wk Moving Avg
119	SPX Member Tech	S&P 500 INDEX	Percentage of Members with Px > 100 Wk Moving Avg
120	SPX Member Tech	S&P 500 INDEX	Percentage of Members with 14 Day RSI Betw 30 & 70
121	SPX Member Tech	S&P 500 INDEX	Percentage of Members with 14 Day RSI > 70
122	SPX Member Tech	S&P 500 INDEX	Percentage of Members with 14 Day RSI < 30

ID	Category	Security name	Filed
123	SPX Member Tech	S&P 500 INDEX	Percentage of Members with New 52 Week Highs
124	SPX Member Tech	S&P 500 INDEX	Percentage of Members with New 52 Week Lows
125	SPX Member Tech	S&P 500 INDEX	Percentage of Members with New 4 Week Highs
126	SPX Member Tech	S&P 500 INDEX	Percentage of Members with New 4 Week Lows
127	SPX Member Tech	S&P 500 INDEX	Pct of Members w/MACD Sell Signal Last 10 Days
128	SPX Member Tech	S&P 500 INDEX	Pct of Members w/MACD Buy Signal Last 10 Days
129	SPX Member Tech	S&P 500 INDEX	Percentage of Members with New 12 Week Highs
130	SPX Member Tech	S&P 500 INDEX	Percentage of Members with New 12 Week Lows
131	SPX Member Tech	S&P 500 INDEX	Percentage of Members with New 8 Week Highs
132	SPX Member Tech	S&P 500 INDEX	Percentage of Members with New 24 Week Highs
133	SPX Member Tech	S&P 500 INDEX	Percentage of Members with New 24 Week Lows
134	SPX Member Tech	S&P 500 INDEX	Percentage of Members with New 8 Week Lows
135	SPX Options and Futures	S&P 500 INDEX	Hist. Call Implied Volatility
136	SPX Options and Futures	S&P 500 INDEX	Put Call Volume Ratio - Current Day
137	SPX Options and Futures	S&P 500 INDEX	Total Option Volume - Current Day
138	SPX Options and Futures	S&P 500 INDEX	Total Call Volume
139	SPX Options and Futures	S&P 500 INDEX	Total Put Volume
140	SPX Options and Futures	S&P 500 INDEX	Total Call Open Interest
141	SPX Options and Futures	S&P 500 INDEX	Total Put Open Interest
142	SPX Options and Futures	S&P 500 INDEX	Total Call Volume Current Day
143	SPX Options and Futures	S&P 500 INDEX	Total Put Volume Current Day
144	SPX Options and Futures	S&P 500 INDEX	Total Call Open Interest Current Day
145	SPX Options and Futures	S&P 500 INDEX	Total Put Open Interest Current Day
146	SPX Options and Futures	S&P 500 INDEX	Total Option Volume - Current Day
147	SPX Options and Futures	Generic 1st 'SP' Future	Aggregate Open Interest
148	SPX Options and Futures	Generic 1st 'SP' Future	Aggregate Volume of Futures Contracts
149	SPX Subindex	S&P 500 Banks Industry Group G	Price Change 1 Day Percent
150	SPX Subindex	S&P 500 Retailing Industry Gro	Price Change 1 Day Percent

ID	Category	Security name	Filed
151	SPX Subindex	S&P 500 Automobiles & Component	Price Change 1 Day Percent
152	SPX Subindex	S&P 500 Transportation Industr	Price Change 1 Day Percent
153	SPX Subindex	S&P 500 Software & Services In	Price Change 1 Day Percent
154	SPX Subindex	S&P 500 Insurance Industry Gro	Price Change 1 Day Percent
155	SPX Subindex	S&P 500 Real Estate Industry G	Price Change 1 Day Percent
156	SPX Subindex	S&P 500 Technology Hardware &	Price Change 1 Day Percent
157	SPX Subindex	S&P 500 Media & Entertainment	Price Change 1 Day Percent
158	SPX Subindex	S&P 500 Household & Personal P	Price Change 1 Day Percent
159	SPX Subindex	S&P 500 Telecommunication Serv	Price Change 1 Day Percent
160	SPX Subindex	S&P 500 Utilities Industry Gro	Price Change 1 Day Percent
161	SPX Subindex	S&P 500 Food Beverage & Tobacc	Price Change 1 Day Percent
162	SPX Subindex	S&P 500 Health Care Equipment	Price Change 1 Day Percent
163	SPX Subindex	S&P 500 Consumer Durables & Ap	Price Change 1 Day Percent
164	SPX Subindex	S&P 500 Pharm Biotech & Life S	Price Change 1 Day Percent
165	SPX Subindex	S&P 500 Energy Industry Group	Price Change 1 Day Percent
166	SPX Subindex	S&P 500 Capital Goods Industry	Price Change 1 Day Percent
167	SPX Subindex	S&P 500 Diversified Financials	Price Change 1 Day Percent
168	SPX Subindex	S&P 500 Food & Staples Retaili	Price Change 1 Day Percent
169	SPX Subindex	S&P 500 Consumer Services Indu	Price Change 1 Day Percent
170	SPX Subindex	S&P 500 Commercial Professiona	Price Change 1 Day Percent
171	SPX Subindex	S&P 500 Materials Industry Gro	Price Change 1 Day Percent
172	SPX Subindex	S&P 500 Consumer Discretionary	Price Change 1 Day Percent
173	SPX Subindex	S&P 500 Consumer Staples Secto	Price Change 1 Day Percent
174	SPX Subindex	S&P 500 Energy Sector GICS Lev	Price Change 1 Day Percent
175	SPX Subindex	S&P 500 Financials Sector GICS	Price Change 1 Day Percent
176	SPX Subindex	S&P 500 Health Care Sector GIC	Price Change 1 Day Percent
177	SPX Subindex	S&P 500 Industrials Sector GIC	Price Change 1 Day Percent
178	SPX Subindex	S&P 500 Information Technology	Price Change 1 Day Percent
179	SPX Subindex	S&P 500 Materials Sector GICS	Price Change 1 Day Percent
180	SPX Subindex	S&P 500 Communication Services	Price Change 1 Day Percent
181	SPX Subindex	S&P 500 Utilities Sector GICS	Price Change 1 Day Percent
182	SPX Tech	S&P 500 INDEX	Volatility 30 Day
183	SPX Tech	S&P 500 INDEX	Volatility 90 Day
184	SPX Tech	S&P 500 INDEX	Volatility 60 Day
185	SPX Tech	S&P 500 INDEX	Volatility 260 Day
186	SPX Tech	S&P 500 INDEX	Volatility 360 Day
187	SPX Tech	S&P 500 INDEX	Volatility 10 Day
188	SPX Tech	S&P 500 INDEX	Volatility 20 Day
189	SPX Tech	S&P 500 INDEX	Volatility 180 Day
190	SPX Tech	S&P 500 INDEX	Volatility 200 Day
191	SPX Tech	S&P 500 INDEX	Volatility 120 Day
192	SPX Tech	S&P 500 INDEX	RSI 3 Day
193	SPX Tech	S&P 500 INDEX	RSI 9 Day
194	SPX Tech	S&P 500 INDEX	RSI 14 Day
195	SPX Tech	S&P 500 INDEX	RSI 30 Day
196	SPX Tech	S&P 500 INDEX	ARMS Daily Index

ID	Category	Security name	Filed
197	SPX Tech	S&P 500 INDEX	ARMS Weekly Index
198	SPX Tech	S&P 500 INDEX	Money Flow Net Non-Block
199	SPX Tech	S&P 500 INDEX	Money Flow Net-Block
200	SPX Tech	S&P 500 INDEX	Dividend Per Share Last Net
201	SPX Tech	S&P 500 INDEX	Volume - Realtime
202	SPX Tech	S&P 500 INDEX	Advance Volumes
203	SPX Tech	S&P 500 INDEX	Decline Volumes
204	SPX Tech	S&P 500 INDEX	Unchanged Volumes
205	SPX Tech	S&P 500 INDEX	Average Volume 5 Day
206	SPX Tech	S&P 500 INDEX	Average Volume 25 Day
207	SPX Tech	S&P 500 INDEX	Moving Average 5 Day
208	SPX Tech	S&P 500 INDEX	Moving Average 10 Day
209	SPX Tech	S&P 500 INDEX	Moving Average 20 Day
210	SPX Tech	S&P 500 INDEX	Moving Average 30 Day
211	SPX Tech	S&P 500 INDEX	Moving Average 50 Day
212	SPX Tech	S&P 500 INDEX	Moving Average 100 Day
213	SPX Tech	S&P 500 INDEX	Moving Average 200 Day
214	SPX Tech	S&P 500 INDEX	Percentage Index Advanced
215	SPX Tech	S&P 500 INDEX	DIFF_MOV_AVG_5D
216	SPX Tech	S&P 500 INDEX	DIFF_MOV_AVG_10D
217	SPX Tech	S&P 500 INDEX	DIFF_MOV_AVG_20D
218	SPX Tech	S&P 500 INDEX	DIFF_MOV_AVG_30D
219	SPX Tech	S&P 500 INDEX	DIFF_MOV_AVG_50D
220	SPX Tech	S&P 500 INDEX	DIFF_MOV_AVG_100D
221	SPX Tech	S&P 500 INDEX	DIFF_MOV_AVG_200D
222	SPX Tech	S&P 500 INDEX	return_5d
223	SPX Tech	S&P 500 INDEX	return_10d
224	SPX Tech	S&P 500 INDEX	return_30d
225	SPX Tech	S&P 500 INDEX	return_60d
226	SPX Tech	S&P 500 INDEX	Max 30 day
227	SPX Tech	S&P 500 INDEX	Days away from 30 Max
228	SPX Tech	S&P 500 INDEX	Min 30 day
229	SPX Tech	S&P 500 INDEX	Day away from 30 Min
230	SPX Tech	S&P 500 INDEX	RSI3d/RSI14d
231	SPX Tech	S&P 500 INDEX	MOV_AVG_5D_20D
232	SPX Tech	S&P 500 INDEX	Price Change 1 Day Percent
233	Vix Tech	Cboe Volatility Index	vixchanged_3d
234	Vix Tech	Cboe Volatility Index	vixchanged_5d
235	Vix Tech	Cboe Volatility Index	vixchanged_10d
236	Vix Tech	Cboe Volatility Index	vixchanged_30d
237	Vix Tech	Cboe Volatility Index	vixchanged_60d
238	Vix Tech	Cboe Volatility Index	vixstd_5d
239	Vix Tech	Cboe Volatility Index	vixstd_10d
240	Vix Tech	Cboe Volatility Index	vixstd_30d
241	Vix Tech	Cboe Volatility Index	vixstd_60d
242	Vix Tech	Cboe Volatility Index	Moving Average 5 Day
243	Vix Tech	Cboe Volatility Index	Moving Average 10 Day

ID	Category	Security name	Filed
244	Vix Tech	Cboe Volatility Index	Moving Average 20 Day
245	Vix Tech	Cboe Volatility Index	Moving Average 30 Day
246	Vix Tech	Cboe Volatility Index	Moving Average 5 Day
247	Vix Tech	Cboe Volatility Index	Moving Average 10 Day
248	Vix Tech	Cboe Volatility Index	Moving Average 20 Day
249	Vix Tech	Cboe Volatility Index	Moving Average 30 Day
250	Vix Tech	Cboe Volatility Index	RSI 3 Day
251	Vix Tech	Cboe Volatility Index	RSI 9 Day
252	Vix Tech	Cboe Volatility Index	RSI 14 Day
253	Vix Tech	Cboe Volatility Index	RSI 30 Day
254	Vix Tech	Cboe Volatility Index	Max 30 day
255	Vix Tech	Cboe Volatility Index	Days away from 30 Max
256	Vix Tech	Cboe Volatility Index	Min 30 day
257	Vix Tech	Cboe Volatility Index	Day away from 30 Min
258	Vix Tech	Cboe Volatility Index	RSI3d/RSI14d
259	Vix Tech	Cboe Volatility Index	MOV_AVG_5D_20D
260	Vix Tech	Cboe Volatility Index	Price Change 1 Day Percent
261	World Equity Index	DOW JONES INDUS. AVG	Price Change 1 Day Percent
262	World Equity Index	NIKKEI 225	Price Change 1 Day Percent
263	World Equity Index	Euro Stoxx 50 Pr	Price Change 1 Day Percent
264	World Equity Index	DAX INDEX	Price Change 1 Day Percent
265	World Equity Index	NASDAQ COMPOSITE	Price Change 1 Day Percent
266	World Equity Index	FTSE 100 INDEX	Price Change 1 Day Percent
267	World Equity Index	STXE 600 (EUR) Pr	Price Change 1 Day Percent
268	World Equity Index	HANG SENG INDEX	Price Change 1 Day Percent
269	World Equity Index	TOPIX INDEX (TOKYO)	Price Change 1 Day Percent
270	World Equity Index	SHANGHAI SE COMPOSITE	Price Change 1 Day Percent
271	World Equity Index	RUSSELL 2000 INDEX	Price Change 1 Day Percent
272	World Equity Index	NASDAQ 100 STOCK INDX	Price Change 1 Day Percent
273	World Equity Index	CAC 40 INDEX	Price Change 1 Day Percent
274	World Equity Index	MSCI World Index	Price Change 1 Day Percent
275	World Equity Index	MSCI Emerging Markets Index	Price Change 1 Day Percent
276	World Equity Index	NSE Nifty 50 Index	Price Change 1 Day Percent
277	World Equity Index	BRAZIL IBOVESPA INDEX	Price Change 1 Day Percent
278	World Equity Index	BarCap US Corp HY YTW - 10 Yea	Price Change 1 Day Percent

Online Appendix II. Details of the research design and list of Algorithms included

OApp II.1 Predictive objective

The objective of this research work is to predict the VIX daily signal for the next day. We choose to predict direction instead of the level of VIX because we take the view from a portfolio manager who is interested in timing the VIX. The forecast will translate to a decision to long or short volatility. Given this objective, forecasting classification is a direct match to this operation problem. We select a supervised machine learning approach, that the algorithm learns from the input data and then uses this learning to predict the VIX's UP or DOWN signals. This is a typical Binominal Classification problem, that can be addressed by several algorithms in Machine Learning.

OApp II.2 Variable selections and data processing

Before applying the ML algorithms to the features described above, we conduct a set of feature engineering processes for the data[22]. The quality of features in the data will directly influence the predictive model's flexibility, simplicity, execution performance, and corresponding results. Especially, certain ML algorithms, such as Tree-based methods, are not sensitive to feature unit and magnitude, but some others are. Therefore, scaling methods (Standardization and Min-Max Scaling) are applied to the data fields, to address the features with highly varying magnitudes, units, and ranges. Such rescaling is done using the training sample (instead of the full sample) for each new model training to avoid looking forward bias.

OApp II.3 Machine Learning algorithm selection

There are recent studies of classifiers in the financial market. For example, in studying stock market prediction, Ballings et al (2015) show the following order

[22] The feature engineering normally covers data cleaning, data scaling and transformation, feature selection, feature enhancement (extraction and enhancement), feature construction and feature learning.

according to AUC (Area Under the Curve): Random Forests, SVM, Kernel Factory, AdaBoost, Neural Networks, K-Nearest Neighbor, and Logistic Regression. Booth, Gerding, and McGroarty (2015) show that an ensemble of random forests improves forecast accuracy by at least 15% compared with the competing models (linear regression, neural networks, and SVM) in the out-of-sample forecast of price impact. For the main analyses of this paper, we include Naïve Bayes (1.NB), Logistic Regression (2.LR), and classic ML such as Decision Tree (3.DT) and Random Forest (4.RF); we also include more advanced methods such as Adaptive boosting (5.AB), Multi-Layer Perceptron (6.MLP) and an Ensemble model (7.Ens) using all of the above. These cover a wide range of model complexity to examine which type of algorithm is better for volatility directional prediction.

Naïve Bayes (NB)

Naïve Bayes classifiers are a family of simple probabilistic classifiers based on Bayes' Theorem with strong (naive) independence assumptions among the features/predictors. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability. The Naive Bayes model is easy to build and particularly useful for large data sets. Along with simplicity, Naive Bayes, in certain circumstances, is known to outperform even highly sophisticated classification methods. In this research work, we use the Naïve Bayes algorithm as a benchmark approach, to compare with other algorithms to demonstrate the model performance.

Logistic Regression (LR)

Logistic Regression is a statistical method to analyze a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable in which the calculated probability is determined by two possible outcomes. The goal of logistic regression is to find the best-fitting model to describe the relationship between the dichotomous characteristic of interest and a set of independent predictors.

In this research work, we use the Naïve Bayes algorithm and Logistic Regression as benchmark approaches, to compare with other algorithms to demonstrate the model performance.

Decision Tree (DT)

A decision tree can be used to build a classification model in the form of a tree structure. It splits the input data set into several smaller subsets, to incrementally develop a tree structure to fit the decision in each node. Therefore, the final result is a tree with decision nodes and leaf nodes. The decision nodes are factors of the decision, and the leaf nodes denoted each decision. A decision node has two or more branches, and each leaf node represents a classification or decision. The topmost decision node in a tree that corresponds to the best predictor is called the root node.

The prediction process using the decision tree is to go through the tree from the root node, to look for the best-fitting factors that led to the decision. It is based on the concept of entropy, which looks at the frequency distribution of decisions and then calculates a logarithm. The process repeats itself, by dividing each decision into sub-conditions for each decision until entropy is zero, and then the best decision is found.

Random Forest (RF)

Random forest is an advanced development on top of the Decision Tree, with randomness to avoid bias and group outcomes based on the most likely positive responses. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. As the number of trees increases and given the law of large numbers, random forests do not suffer from overfitting problems (Breiman, 2001).

Booth, Gerding, and McGroarty (2014) apply an automated trading system based on performance-weighted ensembles of random forests and show better prediction accuracy and higher profitability than other ensemble methods using data on the German stock index DAX.

Adaptive Boosting (AB)

Boosting is a set of algorithms that combines weak learners to form a strong rule. It starts from base learners with a different distribution to generate a new weak prediction rule. With a number of iterations, the boosting algorithm combines weak rules into a single stronger prediction rule.

Among the Boosting algorithms, Adaptive Boosting (AB) is more focused on classification problems and aims to convert a set of weak classifiers into a strong one. It starts with weighted training data to predict the original data set and give equal weight to each observation. If the prediction is incorrect, then the weight is adjusted. This process loops in iterations to continuously add learners until a predefined limit is reached in the number of models, the accuracy, the number of iterations or other user-defined criteria.

Neural Network (MLP)

Neural Network has become popular in recent years in the Machine Learning world, new architectures such as Recurrent Neural Network, and Convolutional Neural Network are used in many areas of image/video recognition, speech recognition, time series analysis etc. The typical neural network consists of units (neurons), arranged in layers. The units are connected, and each connection has a weight associated with it. The whole structure can convert an input vector into an output. Each unit takes an input, applies a nonlinear objective function to it and then passes the output on to the next layer.

The most common model is the Multi-Layer Perceptron (**MLP**), where a unit takes the input data and feeds its output to all the units on the next layer. In the training phase, the weights are tuned to adapt the whole neural network to reach the optimal output of the objective function.

Ensemble model (Ens)

When multiple models are trained, we combine them into a stacking ensemble model, to increase the predictive force of the classifier, and reduce the errors introduced by variance, noise, and bias. First, all the basic models are trained using grid search with cross-validation, to obtain the model structure with a sub-optimal hyperparameter set. Then the ensemble algorithm is applied to

make a final prediction using all the predictions of the individual model, each model can be weighted based on their individual performance. This approach typically yields performance better than any single one of the trained models. In this paper, the ensemble model is built with six algorithms (Naïve Bayes, Logistic Regression, Adaptive Boosting, Decision Tree, Random Forest, and Multi-Layer Perceptron) that are described above. The "ensemble" approach has been shown to produce better predictions than any individual AI technique (Rasekhschaffe, and Jones 2019).

Online Appendix III. Support Vector Machine

We study the Support Vector Machine (SVM) for VIX prediction in this appendix. Table OA1 reports the performance of SVM compared with other models. The overall accuracy rate of SVM is relatively low compared to other ML algorithms. It is comparable with the simple HAR model. However, it has an extremely low timing ratio. In other words, it produces a highly unbalanced prediction. It has a 'down' prediction bias in its prediction. Panel B shows that this one-sided prediction produces a negative overall return which is also observed in a short-only strategy. Overall, our experiment shows that SVM produces a less effective accuracy rate due to its unbalanced prediction despite the use of a balanced sample. Previous studies find supportive evidence to SVM models is typically small system with fewer inputs. Our system with 278 variables takes much longer to run and it seems to produce a 'corner solution'.

<Insert Table OA1>

Online Appendix IV. Examples of practical applications

In this section, we present two examples of practical investment applications. We first study strategies of trading VIX futures with the model signal and compare them with the short-only strategy. We then present some limited evidence of trading the contract for difference (CFD).

Trading VIX futures

We consider a direct application of our forecast to trade VIX derivatives[23]. Among them, VIX futures are relatively liquid with lower trading costs (a narrow spread). However, since VIX and VIX futures do not have the traditional cost of carrying relationship. Their movements need not be perfectly correlated. Therefore, how useful the VIX prediction is for trading VIX futures is an empirical question.

We study this by reconstructing a continued series of VIX futures returns. This is done by using the front-month contract and switching to the next one when the current one expires. This is similar to the '1st generic series' in Bloomberg called UX1. However, using the UX1 series to calculate return will be misleading during the switch. We carefully identify the day of switching and calculate the return using the new contract before the day of switching.

Three key features of this continued VIX future time series are worth noting. First, the daily correlation between the return of the nearest future contract and VIX is about 89%. Second, simple regression analysis shows that the VIX futures return is 0.6 of the VIX return and the r-squared of the regression is 63%. Third, this time series (cumulated return) is declining with time. It is a well-known fact that VIX futures are traded at a premium to VIX in normal conditions (Rhoads, 2011, Chapter 11). In normal market conditions, the term structure of VIX is similar to a normal yield curve upward sloping.

[23] VIX options contracts are delivered at the VIX closing, therefore, they are the only product that is directly linked to the spot VIX which are the target that our model predicts. However, this product is very illiquid with a premium of more than 10% when one tries to trade near the money contracts for the weekly VIX. It will be not worth further investigation with our average daily level of return at less than 1%. Alternative options strategies may be possible to exploit the predictability of our model. We leave this to future research.

Therefore, there is a short VIX futures premium that inspired the creation of Short VIX ETF such as the "ProShares Short VIX Short-Term Futures". In the end, CBOE creates the CBOE VIX Premium Strategy Index (VPD) for benchmarking such investment opportunities[24]. The existence of such a premium (a 'short bias') reduces the correlation between the VIX spot and VIX futures.

Bearing the above facts in mind, we explore the application of our VIX signal to VIX futures in two ways. First, it is a direct application without modification and second, it considers the short bias.

[24] It reports the return of a strategy consistently selling front month VIX futures combined with a money market account.
https://ww2.cboe.com/micro/vpd/vixpremiumindexvpdvpn.pdf

Table Table OA2 reports the summary statistics of applying the VIX signal directly on the nearest month VIX futures contract. As expected, a reduction in accuracy rate is observed across the board (a drop ranging from 1.7% to 4.4%). While most of the before-cost return is positive, they are much smaller in magnitude than expected. This is caused both by the drop in accuracy and the fact that the magnitude of the VIX futures movement is on average only 60% of the VIX. For example, the best-performing strategy AB has its return reduced to 1/3 of the VIX return. For transaction costs, the VIX futures bid-ask spread is on average 41 basis points[25]. None of the signals would earn a positive return should we rebalance this *every day*. However, we do not need to rebalance daily as predictions can be persistent. The turnover ratio suggests that we only need to trade 23% to 38% of the days. When we are taking into consideration the spread dynamically by applying the spread only when the prediction direction changes, most of the strategies still have a positive return. However, they are of a much smaller magnitude. For the AB model, it has an average return of 15 basis points daily which annualized to 38% annually. While this is nowhere near 2.25 times in the VIX application, a 38% return on average in 11 years would be an excellent achievement even for historically successful star investors. Overall, this type of application reduces performance from both the low correlation between the VIX and future and the low relative size of the return.

<Insert Table OA2>

We now turn to the consideration of the shorting volatility strategy. We consider using our signal with the knowledge of the short-bias in the futures data would be able to beat the short-only strategy. To this end, we consider a modified trading strategy that avoids going against this short bias in the VIX futures. Specifically, this would affect our trading in the long leg. When our prediction is UP but the VIX futures are currently higher than the VIX spot, this will put downward pressure on the VIX futures contract. If we were to long the VIX futures contract following our model prediction for the VIX spot, this would go against the general expected negative VIX futures return given that it

[25] Calculated by the bid and ask price of the contracts during the sample period.

is likely to converge to the VIX spot. We, therefore, choose to "sit out" on this type of prediction day.

Table OA3 shows that avoiding the 'conflicting' days significantly improved the correct ratio. It brings them back to the levels that are comparable to those we saw in the simulated strategy of the VIX spot in Section 0. The number of investment days shows that most of the model sits out less than half of the trading days except for NB. For comparison, we include the short-only (SO) strategy as a benchmark. We have four key observations from the results. First, the short-only (SO) strategy is one that hard to beat especially when the rebalancing costs taking into consideration. It produces a 22 basis points daily return with a Sharpe ratio of 0.57. As expected, this one-way bet produces zero market timing.

Second, most of the models' average accuracies improve by sitting out those conflicting days but not by too much. The most consistently performed models are LR, AB, ENS and DT judging by their *t* value before costs. For the magnitude of the return, all models have a higher before-cost return than SO. Third, taking into consideration the turnover and transaction costs of the strategy dynamically, AB is the only model that can convincingly beat the SO strategy. It has about a 50% higher daily average return. Importantly it has a much higher Sharpe ratio (0.93 vs 0.57) and a lower yearly MDD. Finally, the LR model performs reasonably well with measures slightly better than SO in every aspect. These results give confidence about the financial and economic variables in our studies that are relevant to the VIX forecast even in a 'linear' model. Overall, this analysis demonstrates the economic significance of our forecast.

<Insert Table OA3>

Trading VIX CFD (spread betting)

There is another way to trade VIX which is the contract for difference (CFD) or spread betting from platforms such as IG.com[26]. However, the IG only provides

[26] https://www.ig.com/uk/indices/markets-indices/volatility-index accessed March 2021.

limited historical data. Their daily data is recorded at 5 am UK time every day which is not in the normal VIX calculation trading hour. This daily close return has a very low correlation with VIX. To obtain prices that are within the VIX calculation hour, we obtain the hourly data (one year's worth of data) in March 2021. We test our trading strategy from 11 March 2020 to the end of 2020 (197 days) trading near the VIX closing which is 3 pm Central Time[27].

In Table OA4 we report the statistics for the IG application. We also report the strategy return for applying to the VIX data for comparison. First, for accuracy, there is a drop in the accuracy when applying to IG in most models as expected except for the first two basic models and the ensemble model. In general, the changes are between 1% to 3%. Second, when looking at the return before costs, the IG returns are uniformly lower while maintaining similar patterns as applying to VIX. The decision tree families model continued to do especially well. Third, the main transaction costs in trading CFD are the spread. During this period, the spread is relatively high since the Covid 19 pandemic increases the volatility of the VIX. It is 0.16 times the price level for each contract[28]. We add these costs to the trading when there is a change of direction in the predictions. On average the turnover ratio is between 14% to 46%. The decision tree's models (DT, RF, and AB) continued to produce good returns after costs. Especially the DT model. We also report the cumulative return during the 197 trading days (end on 31 Dec 2020). The AB doubles the investment during this period without taking into consideration of leverage[29]. However, there is a risk attached to these strategies. Only the DT return is statistically significantly different from zero suggesting other returns are

[27] The VIX Index is calculated between 2:15 a.m. CT and 8:15 a.m. CT and between 8:30 a.m. CT and 3:15 p.m. CT. We will not be able to have meaning full length of historical data for 15-minute frequency.

[28] Note that this spread is in absolute price value that will deducted from the transaction price. Therefore, the percentage transaction costs would be changing. For example, if VIX is around 20, then the percentage spread is 0.8%. This percentage spread will be halved (0.4%) if VIX increase to 40.

[29] CFD or spread betting are leverage product by design. The normal margin is 20% which produce 1:5 leverage ratio. Leverage is a double-edged sword. It is particularly harmful for VIX trading when VIX spike and the trade is on the wrong side. One can de-leverage the position even trading with this product by taking a proportionally smaller position to the capital that one willing to invest. For example, only use 20% of your capital in this strategy while calculating return with the full capital will produce an unleveraged return. All the return report here does not consider leverage. They are just the normal long-short return assuming 100% investment of capital.

volatile around zero. The maximum drawdown can be as high as 58% for the DT strategy. It is also important to note that this testing period is short and starts right after the pandemic spike. Finally, the short-only strategy (SO) can produce high accuracy of directional prediction, but it often misses out on the relatively larger size movement and the overall return is similar to NB during this period. This confirms the value of economic data and modelling in this context.

<Insert Table OA4>

Table OA1. SVM performance

This table reports the performance of SVM comparing with other MLs. 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 (true positive ratio + true negative ratio) -1. Detail of the models is in section App.3. 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.

Panel A. Accuracy and Timing

	Correct ratio	Information coefficient		Timing ratio		Positive hit	Negative hit
Models	**Mean**	**Mean**	**t**	**Mean**	**t**	**Mean**	**Mean**
1.NB	0.487	-0.0262	-1.36	0.020	1.35	0.770	0.250
2.LR	0.552	0.1046	4.19***	0.097	4.94***	0.480	0.620
3.DT	0.582	0.1634	6.93***	0.126	5.25***	0.350	0.780
4.RF	0.560	0.1194	10.64***	0.122	9.60***	0.570	0.550
5.AB	0.569	0.1383	8.01***	0.122	7.38***	0.480	0.650
6.MLP	0.530	0.0595	2.33**	0.057	2.61**	0.580	0.480
7.ENS	0.566	0.1319	7.19***	0.119	6.41***	0.490	0.630
HAR	0.525	0.0502	2.31**	0.072	4.02***	0.680	0.390
SO	0.546	0.0928	6.78***	0.000.		0.000	1.000
SVM	0.523	0.0452	2.80**	0.002	0.28	0.220	0.790

Panel B Simulated return applied to VIX.

	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***
SVM	-0.0014	-1.80	-0.54	-75%	-17.23***

Table OA2. Forecast accuracy and return applying to VIX futures

This table reports the performance of a strategy trading the out-of-sample VIX prediction signal on the nearest month VIX futures contract. The correct ratio is the proportion of prediction days that have a correct prediction. For the strategy returns, we report the mean daily return and the Sharpe ratio for both before and after costs. The costs are measured by the bid-ask spread of the contract. Costs are applied when there is a change in the trading direction. Turnover measures the proportion of the days that need to rebalance due to the change of trading direction. For the after-cost return, we also report the average maximum annual percentage drawdown (MDD) and the 'maximum' of the MDD in the 11 years. ***, **, and * indicate statistical significance at 1%, 5% and 10% level respectively.

	Correct ratio	Before Cost Return			After Cost Return			Yearly MDD		Investment days	turnover
Model	**Mean**	**Mean**	**t**	**Sharpe**	**Mean**	**t**	**Sharpe**	**Mean**	**Max**	**Mean**	**Mean**
1.NB	0.445	-0.0023	-2.86**	-0.86	-0.0031	-3.79***	-1.14	-0.73	-1.23	233	0.23
2.LR	0.535	0.0023	3.78***	1.14	0.0008	1.28	0.39	-0.42	-1.32	233	0.37
3.DT	0.555	0.0023	2.29**	0.69	0.0011	1.17	0.35	-0.52	-1.36	233	0.31
4.RF	0.515	0.0014	1.53	0.46	0.0000	0.04	0.01	-0.49	-1.18	233	0.33
5.AB	0.526	0.0029	2.56**	0.77	0.0015	1.48	0.44	-0.43	-0.72	233	0.38
6.MLP	0.508	0.0005	0.56	0.17	-0.0008	-0.76	-0.23	-0.53	-1.24	233	0.30
7.ENS	0.525	0.0020	2.46**	0.74	0.0006	0.73	0.22	-0.58	-1.43	233	0.33
HAR	0.470	-0.0006	-0.70	-0.21	-0.0014	-1.48	-0.45	-0.62	-1.15	233	0.19

Table OA3. Augmented VIX futures strategy: sitting out on conflicting days to avoid shorting pressures

This table reports an augmented VIX futures trading strategy. The strategy trades the out-of-sample VIX prediction signal on the nearest month VIX futures contract except on the days that the model predicts an up signal while the VIX futures contract is trading higher than the VIX spot value. The correct ratio is the proportion of prediction days that have a correct prediction. For the strategy returns, we report the mean daily return and the Sharpe ratio for both before and after costs. The costs are measured by the bid-ask spread of the contract. Costs are applied when there is a change in the trading direction. Turnover measures the proportion of the days that need to rebalance due to the change of trading direction. For the after-cost return, we also report the average maximum annual percentage drawdown (MDD) and the 'maximum' of the MDD in the 11 years. ***, **, and * indicate statistical significance at 1%, 5% and 10% level respectively.

	Correct ratio	Timing ratio	Before Cost Return			After Cost Return			Yearly MDD		Investment days	turnover
Model	Mean	Mean	Mean	t	Sharpe	Mean	t	Sharpe	Mean	Min	Mean	Mean
1.NB	0.529	0.005	-0.0001	-0.03	-0.01	-0.0023	-0.76	-0.23	-0.48	-1.21	81	0.60
2.LR	0.593	0.026	0.0040	4.04***	1.22	0.0024	2.24**	0.67	-0.32	-1.28	151	0.36
3.DT	0.599	0.002	0.0031	2.15**	0.65	0.0021	1.52	0.46	-0.47	-1.29	174	0.24
4.RF	0.594	0.006	0.0035	1.99*	0.60	0.0018	1.10	0.33	-0.36	-1.18	128	0.36
5.AB	0.593	0.016	0.0050	3.95***	1.19	0.0036	3.08**	0.93	-0.33	-0.62	149	0.37
6.MLP	0.577	0.014	0.0032	1.98*	0.60	0.0014	0.84	0.25	-0.34	-1.21	124	0.41
7.ENS	0.592	0.006	0.0036	2.60**	0.79	0.0022	1.51	0.45	-0.45	-1.34	145	0.32
HAR	0.562	0.004	0.0019	0.74	0.22	0.0004	0.15	0.04	-0.48	-1.19	98	0.36
SO	0.594	0.000	0.0022	1.88*	0.57	0.0022	1.88*	0.57	-0.41	-1.24	233	0.00

Table OA4. Trading CFD with VIX forecasts

This table reports the performance of the models applying the VIX signals to the spread betting data from IG.com between 11 Mar 2020- 31 Dec 2020. The hourly data is downloaded, and the trade is executed at 3 pm US central time which is 15 minutes before the VIX closing. It reports the statistics for analyses using both the IG price and the VIX hourly data which is downloaded from (Refinitiv Tick History). It reports the correct ratio and the mean before costs return. For IG, the after-costs mean return, cumulated return (sum, no compounding), maximum drawdown and turnover of the strategy are reported. N report the number of days. ***, **, and * indicate statistical significance at 1%, 5% and 10% level respectively.

	VIX	IG	VIX_r	IG_r	IG_r_after_cost				
Models	**correct**	**correct**	**Mean**	**Mean**	**Mean**	**Cumulated**	**MDD**	**turnover**	**N**
1.NB	0.5	0.5	-0.0033	-0.0029	-0.0036	-0.65	-0.67	0.14	197
2.LR	0.44	0.47	0.0018	-0.0010	-0.0021	-0.51	-0.53	0.2	197
3.DT	0.65	0.63	0.0185***	0.0102***	0.0081**	2.59	-0.58	0.37	197
4.RF	0.55	0.53	0.0123**	0.0080**	0.0058	1.34	-0.22	0.41	197
5.AB	0.59	0.58	0.0189***	0.0077*	0.0052	1.03	-0.37	0.45	197
6.MLP	0.57	0.54	0.0083	0.0007	-0.0009	-0.39	-0.68	0.29	197
7.ENS	0.54	0.55	0.0116**	0.0043	0.0021	0.11	-0.65	0.38	197
SO	0.59	0.57	0.0035	0.0023	0.0023	0.14	-0.52	0.01	197