Master's Thesis

Forecasting the CBOE Volatility Index: Application of Random Forest and its Variable Importance Measure

Byung Yeon Kim

The Graduate School

Sungkyunkwan University

Department of Economics

Forecasting the CBOE Volatility Index: Application of Random Forest and its Variable Importance Measure

Byung Yeon Kim

A Master's Thesis Submitted to the Department of Economics and the Graduate School of Sungkyunkwan University in partial fulfillment of the requirements for the degree of Master of Arts in Economics

October 2020

Supervised by

Heejoon Han

Major Advisor

Contents

1. INTRODUCTION 1

1.1. Motivation 1

1.2. Forecasting the VIX Index 3

2. RANDOM FOREST AND VARIABLE IMPORTANCE 5

2.1. Classification and Regression Trees (CART) 5

2.2. Random Forest and Permutation Importance 7

2.3. The Boruta Algorithm 9

3. DATA AND METHODOLOGY 13

3.1. Data 13

3.2. Methodology 15

4. RESULTS 17

4.1. Variable Ranking by the Boruta Algorithm 17

4.2. The Optimal Number of Variables 19

4.3. Forecasting Results 20

4.4. Robustness Check 25

5. CONCLUSION 27

Abstract

Forecasting the CBOE Volatility Index: Application of Random Forest and its Variable Importance Measure

The CBOE volatility index (VIX) stands as a representative barometer of the overall sentiment and volatility of the financial market. This paper seeks to apply random forest and its variable importance measure to forecasting the VIX index. Compared to the previous literature which finds it difficult to beat the pure HAR process in forecasting the VIX index due to its persistent nature, random forest could produce forecasts that are significantly more accurate than the HAR and augmented HAR models for multi-days forecasting horizons. Also, the superior predictability of random forest compared to the linear models becomes more evident for the longer forecasting horizons. The forecasting accuracy of random forest could further be improved by systematically selecting the optimal number of the most important variables from a dataset of 298 macro-finance variables, using the Boruta algorithm based on random forest's variable importance measure.

Keywords : Random Forest, Variable Importance, Machine Learning, VIX Index, Volatility Forecasting

1. INTRODUCTION

1.1 Motivation

The implied volatility index of the Chicago Board Options Exchange (CBOE), commonly known as the VIX index, represents the market's estimate of the future volatility of the S&P 500 over the next 30 calendar days. It is derived from the bid/ask quotes of options on the S&P 500 index and is disseminated on a real-time basis. Being calculated directly from the option prices rather than being solved out of an option pricing formula like the Black-Scholes, the VIX index is free from the measurement errors that were present in the previous implied volatility measures.

Today, the VIX index is receiving much attention in the financial market. Not only is it widely traded in the form of VIX futures for hedging or speculative purposes, but also it is being acknowledged as the world's leading barometer of investor sentiment and market volatility. Thus, accurate forecasts of the VIX index in the short and long term can provide crucial information to the participants in the financial market.

The recent advances in machine learning (ML) methods and the accessibility to "big" datasets have opened opportunities to approach the problem of forecasting economic time series in a novel way. While traditional econometric applications are centered around the parameter estimation of $\hat{\beta}$, ML methods

revolve around problem of prediction – producing predictions of $\hat{y}$ from x. Where traditional econometric models rely on careful assumptions about the underlying data-generating-process, ML methods seeks to discover complex structures that are not specified in advance. It manages to fit complex and very flexible functional forms to the data without simply overfitting, and to produce relatively accurate out-of-sample predictions (Mullainathan, 2017).

Medeiros et al. (2019) is one of the recent researches in the literature that has highlighted the benefits of applying ML methods to economic time series forecasting. It applies a wide range of ML models to forecasting US inflation and finds that the forecasts of traditional benchmark models can be improved significantly with ML methods combined with new high-dimensional datasets. Especially, it reports that a particular model, random forest of Breiman (2001), outperforms all other models consistently due to its ability to catch nonlinearities and its variable selection mechanism. Moreover, it reports that the superiority of random forest becomes more evident in settings where the forecasting horizon becomes longer and during the periods when the volatility is higher.

This research seeks to discover how ML methods can provide benefits to forecasting the VIX index, especially with a focus on the random forest and its variable selection mechanism through variable importance measures.

1.2 Forecasting the VIX Index

While there are numerous research topics around implied volatility and the VIX index, the literature agrees on that there is only a handful of research that focuses directly on forecasting the VIX index.

Among the first of these is Ahoniemi (2006) which compares predictive abilities of ARIMAX−GARCH, ARFIMA and Probit models in daily forecasting of VIX. Considering eight financial and macroeconomic variables, it finds addition of exogenous regressors enhances the forecasting performance, while addition of GARCH terms do not improve forecast accuracy.

Degiannakis (2008) considers realized and conditional volatility of the S&P 500 as exogenous variables to model VIX in an ARFIMA model. However, it concludes that the VIX index is hard to forecast and does not seem to be closely connected to the volatility of the underlying index.

Konstantinidi et al. (2008) models seven different implied volatility indices including VIX in a multivariate VAR framework, confirming the presence of implied volatility spillover between various markets. However, it does succeed in deriving significantly improved forecasts.

Fernandes et al. (2014) applies a heterogeneous autoregressive (HAR) model coupled with neural network approximation to capture non−linearities for forecasting VIX. However, they find little evidence of nonlinearity and concludes that it is very hard to beat the pure HAR process due to the very persistent nature of the VIX index.

Conversely, Psaradellis et al. (2016) finds significant evidence of strong non-linearity in VIX by employing a HAR process combined with support vector regression model, improving the results of the one-day-ahead forecasts of pure HAR model. No exogenous variables are considered under the rationale that their forecasting performances were poor in the previous literature.

Ballestra et al. (2019) focus on the directional forecast of VIX Futures instead of the VIX index and uses a feed-forward neural network model with non-lagged explanatory variables that are available only a few hours before the opening of the CBOE. They find that the neural network model with only one most recent exogenous variable is the superior model, with mean directional accuracy of 65.8%.

Most of the literature focuses solely on one-day-ahead forecasts of VIX, while suggesting the multi-day-ahead forecast problem as topic of future research. Forecasting VIX on a longer horizon can be a significant matter in several aspects. For an investor who adjusts his portfolio including VIX futures on a multi-day basis considering trading costs, multi-day-ahead forecast may be more useful than a one-day-ahead one. For a market participant looking for clues about the future volatility and direction of the overall market, an accurate multi-day-ahead forecast of VIX can provide significant information. It is of note that this paper, unlike most others in the literature, has its focus on multi-day-ahead forecast of VIX up to 22 trading days.

Another innovation of the paper is that while most of the literature on VIX forecasting considers only a handful of or no exogenous variables, this research utilizes a high-dimensional dataset of 298 macro-finance variables. Especially using the Boruta algorithm, it attempts to systematically select the optimal number of the most important variables for random forest. To the best of my knowledge, it is the first in the literature to apply algorithmic variable selection of macro-finance variables in a random forest setting for forecasting VIX.

2. RANDOM FOREST AND VARIABLE IMPORTANCE

2.1 Classification and Regression Trees (CART)

Random forest has its roots in classification and regression trees (CART). Introduced by Breiman et al. (1984), it is a simple model which partitions the predictor space into rectangles using binary splits and then uses the splits to determine the outcome prediction. That is, it divides the set of possible values of the predictors $X_1, X_2, \ldots, X_P$ into J distinct and non-overlapping regions, $R_1, R_2, \ldots, R_J$. For every observation that falls into region R_J, same prediction is made which is simply the mean of the response values for the training

observations in R_J. For a regression tree, the objective is to find the partition $R_1, R_2, \ldots, R_J$ such that RSS given by

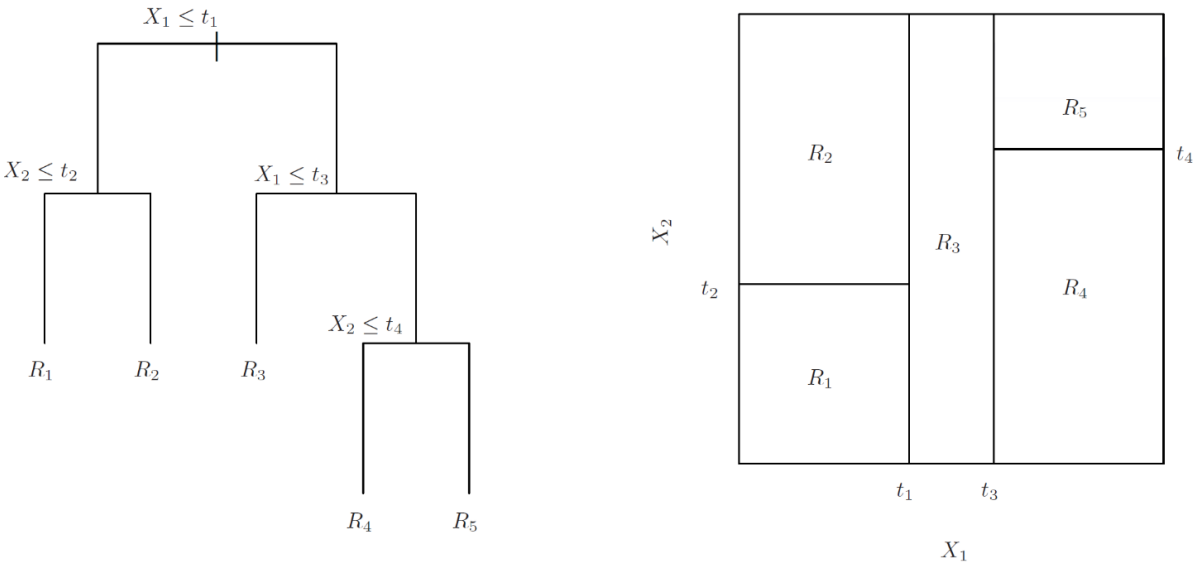
$$\sum_{j=1}^{J} \sum_{i \in Rj} (y_i - \hat{y}_{R_j})^2 \qquad (1)$$

is minimized.

Apparently, it becomes computationally infeasible to consider every possible partition of the predictor space. Thus, a top−down approach known as the recursive binary splitting is utilized. The tree diagram in the left of Figure 1 illustrates a widely used example from Hastie et al. (2001) in which a tree model is grown in a regression setting with two predictors – X_1 and X_2. On the top node (or split) of the tree, the predictor space is partitioned into two regions at $X_1 = t_1$. Then, the region to the left of $X_1 = t_1$ is partitioned at $X_2 = t_2$ and the region to the right is partitioned at $X_1 = t_3$. Finally, the region to the right of $X_1 = t_3$ is partitioned at $X_2 = t_4$. At each node of the tree, the best split is determined such that the decrease in RSS due to the particular split is maximized. It is a greedy approach in that each split considers only the best one at that particular step, instead of looking ahead to consider the future steps. The resulting partition of the predictor space is illustrated in the right diagram of Figure 1, where the five regions (or rectangles) $R_1, \ldots, R_5$ correspond to the five terminal nodes in the tree diagram.

An obvious question one faces when growing a tree model is on how large the

Figure 1: Example of a Regression Tree



tree should be grown. A very large tree could easily overfit the data, whereas a very small tree could miss out on important structures underlying in the data. A widely used strategy to determine the optimal tree size is what is known as the cost−complexity pruning. The idea is to grow a sufficiently large tree, and then prune the tree back to obtain a subtree that minimizes a cost−complexity criterion that penalizes the size of the tree model.

2.2 Random Forests and Permutation Importance

While having low model bias, a single tree model is typically known to be less competitive with the best ML methods in terms of prediction accuracy due to its high variance. Introduced by Breiman (2001), random forest seeks to reduce the variance of trees through a bootstrap aggregation (or bagging) approach.

Thus, the idea is to average many noisy but approximately unbiased trees to obtain stability while taking advantage of the preferable qualities of tree models.

Random forest (RF) is an ensemble of few hundreds to thousands of unpruned trees, each trained on a bootstrap sample of the original data. When building a tree from a bootstrapped sample, RF uses m randomly selected input variables at each split.[1] This random selection of potential predictors to be selected ensures that the trees in the forest are decorrelated to each other. For a regression problem, the prediction of RF for a new test point x is defined as

$$\hat{f}_{rf}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \qquad (2)$$

where B is the number of trees in the whole forest and $T_b(x)$ corresponds to the prediction from the b^{th} tree.

A desirable by-product of the bootstrap sampling process of RF is the presence of out-of-bag (OOB) samples, or the observations that are left out from each bootstrap sampling. These OOB samples can be utilized to measure the importance of each input variables.

When the b^{th} tree is grown, the OOB samples are run down the tree to calculate the OOB mean squared error (MSE):

[1] A typical choice of m with $\mathrm{m} = \sqrt{p}$ for classification and m = p/3 for regression is known to perform well in most cases.

$$OOBMSE_b = \frac{1}{n_{OOB,b}} \sum_{i=1:i\in OOB_b}^{n} (y_i - \hat{y}_{i,t})^2 \quad (3)$$

where $n_{OOB,b}$ denotes the number of observations in the b^{th} OOB sample. Then, the values of the j^{th} variable X_j are randomly permuted in the OOB data, and the permuted OOB MSE is calculated for the j^{th} variable:

$$OOBMSE_b(X_j\ permuted) = \frac{1}{n_{OOB,b}} \sum_{i=1:i\in OOB_b}^{n} (y_i - \hat{y}_{i,t}(X_j\ permuted))^2 \quad (4)$$

If X_j does not have predictive value for the given tree, random permutation of X_j should make small difference to the OOB MSE. On the other hand, if X_j is used as an important variable within the tree, there should be significant decrease in OOB MSE when X_j is randomly permuted. Thus, the decrease in accuracy due to this permutation averaged over all trees is used as a measure of the importance of X_j.

$$\frac{1}{B}\sum_{b=1}^{B}(OOBMSE_b - OOBMSE_b(X_j\ permuted)) \quad (5)$$

This measure of variable importance in RF is known as the *permutation importance*.

2.3 The Boruta Algorithm

One of the strengths of machine learning methods such as RF is their ability

to handle datasets with high-dimensional covariates. However, many machine learning algorithms exhibit a decrease of accuracy when the number of variables is significantly higher than optimal (Kohavi et al. 1997). Thus, when given a high-dimensional dataset, it is often an important matter to distinguish and select out the most important variables; not only for technical efficiency, but also to enhance accuracy in solving the problem that is in question. The variable importance measure of RF described in the previous section can be used as a benchmark for such variable selection procedures.

Until recently, various methodologies have been developed for variable selection using RF variable importance measures. The development has especially been vigorous in the bioinformatics and related fields, i.e. for identifying the important genetic variables for predicting certain dieses status such as cancer. However, there yet seems to be no consensus on a single outperforming variable selection methodology in a RF setting.

Speiser et al. (2019) compares the performance of 13 different RF variable selection procedures that have been developed. It reports the OOB errors as well as the computation time of the different methodologies when they are applied to 311 different datasets. In the research, the *Boruta algorithm* of Kursa et al. (2010) is reported to be one of the better performing procedures overall in terms of OOB errors, and especially a preferrable one in a high-dimensional setting with over 50 predictors.

The Boruta is a wrapper algorithm built around RF that provides a stable

selection of the important variables from the dataset. The Z-score, which is derived for each variable by dividing the permutation importance measure by its standard deviation, is used as the measure of selection. Moreover, it extends the dataset by adding variables that are random by design. For each variable in the dataset, it creates a 'shadow attribute' which is obtained by shuffling values of the original variable.

In detail, the Boruta algorithm consists of the following steps:

1. Extend the dataset by adding copies of all variables.

2. Shuffle the added variables to remove their correlations with the response. (Shadow attributes)

3. Run a random forest on the extended dataset and gather the Z scores computed.

4. Find the maximum Z score among shadow attributes (MZSA), and then assign a hit to every variable that scored better than MZSA.

5. For each variable with undetermined importance perform a two-sided test of equality with the MZSA.

6. Deem the variables which have importance significantly lower than MZSA as 'unimportant' and permanently remove them from the dataset.

7. Deem the variables which have importance significantly higher than MZSA as 'important'.

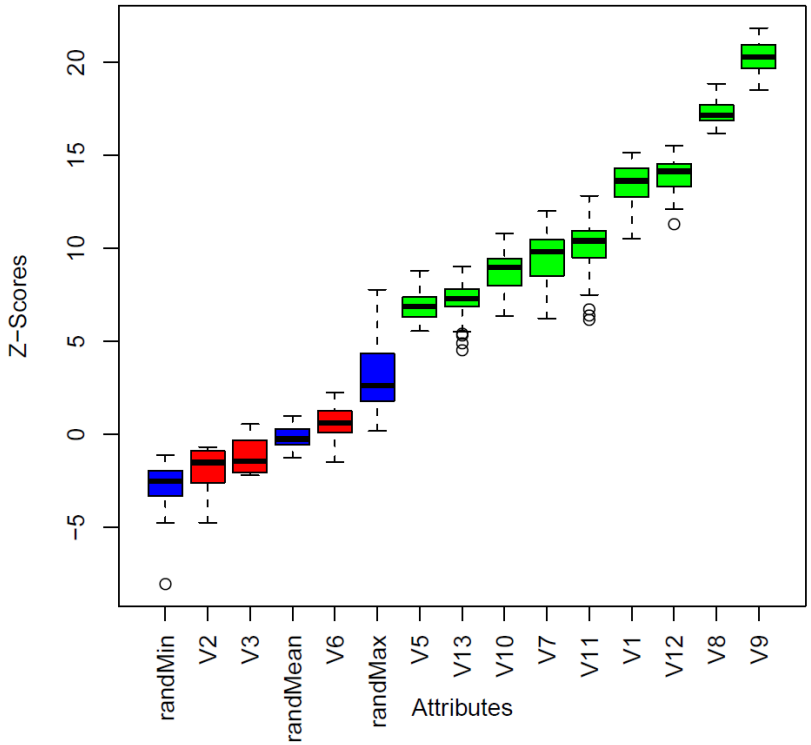
8. Remove all shadow attributes.

9. Repeat the procedure until the importance is assigned for all the variables,

or the algorithm has reached the previously set limit of the random forest runs.

Through an iterative process of eliminating variables deemed unimportant, the Boruta algorithm can deal with both the fluctuating nature of the RF variable importance measure as well as the interactions between the variables. Figure 2 shows an example of a Boruta result plot for a dataset with 13 variables. As can be seen, the distribution of Z−scores from the iterations are displayed and the ranking of the relative importance between variables can be derived.

Figure 2: Example of a Boruta Result Plot for Dataset with 13 Variables



3. DATA AND METHODOLOGY

3.1 Data

The sample period considered in the research starts from April 5, 1990 up to January 15, 2013, matching that of the data used in Fernandes et al (2014). The sample period includes a total of 5,740 daily observations of the VIX index and all exogenous variables. The dataset is divided into the training and the test period. That is, the models are trained or fitted on the first 2,500 daily observations (April 5, 1990 – February 28, 2000) and tested on the rest of 3,240 daily observations (February 29, 2000 – January 15, 2013).

Table 1 shows the descriptive statistics for the logarithmic of the VIX index throughout the whole sample period. The mean, median, minimum, maximum, standard deviation, skewness, kurtosis, the p-values from Augmented Dickey-Fuller (ADF), Phillips-Perron (PP) tests and the test statistic of KPSS test are reported.

This research considers 298 other macro-finance variables as explanatory variable for forecasting the VIX index, which are as listed in the following: the k-day continuously compounded returns on the S&P 500 index (k=1,5,10,22) and the first difference of the logarithm of the volume of the S&P 500 index; the k-day continuously compounded returns on the crude oil futures contract; the first difference of the logarithm of US dollar foreign exchange value on

seven currencies (Australian dollar, Canadian dollar, Swiss frac, euro, British sterling pound, Japanese yen and Swedish kroner) and a trade-weighted average of the above foreign exchange values; the difference in returns between Moody's seasoned Baa rated corporate bonds and Aaa rated corporate bond (credit spread); the difference between 10-year and 3-month treasury constant maturity rates (term-spread); the first difference of the logarithm of 10 other stock indices (NASDAQ-100, Dow Jones Industrial Average, FTSE ALL-Share index, FTSE-100 index, DAX Performance index, Swiss Market index, Nikkei 225, KOSPI index, Hang Seng index, and the BSE Sensex index); the first difference of the logarithm of world gold price; and the daily returns of the individual S&P 500 composites available since 1990 (266 return series).

Table 1: Descriptive Statistics for Logarithm of VIX
(April 5, 1990 - January 15, 2013)

Mean	2.951
Median	2.931
Minimum	2.231
Maximum	4.393
Standard Deviation	0.349
Skewness	0.547
Kurtosis	3.274
Jarque-Bera	0.000
ADF	0.000
PP	0.000
KPSS	0.064

Notes: The critical values of the KPSS test are 0.119, 0.146 and 0.216 at the 10%, 5% and 1% level, respectively.

All data were retrieved from Thomson Reuters Datastream and Federal Reserve Economic Data (FRED).

3.2 Methodology

3.2.1 Out-of-sample forecasting

For direct comparison of results, the same forecasting methodology is followed as described in Fernandes et al. (2014). Forecasts are made from a rolling window of fixed size. For each model, a rolling window of 2,500 time-series observations is used to estimate the model, and out-of-sample forecast evaluation is performed on the remainder of the dataset (3,240 observations). Direct forecasts are made with no consideration into forecasting the covariates.

Four different forecasting horizons of k-day(s) (k=1, 5, 10, 22) are considered. k-day(s)-ahead forecasts are made and mean squared error (MSE) and mean absolute error (MAE) are calculated for each model and forecasting horizon. The MSE and MAE of the random forest are compared to the benchmark models.

3.2.2 Benchmark Models

The benchmark on which to compare the results are the models whose performance is reported in Fernandes et al. (2014). Namely, they are the random walk (RW) model, Autoregressive model with exogenous variables (ARX), heterogeneous autoregressive (HAR) model of Corsi (2009), HAR

model with exogeneous variables (HARX), and HARX model with neural network specification (NNHARX). The model specifications follow that as described in Fernandes et al. (2014). For the models including exogeneous variables, the 14 variables used in Fernandes et al. (2014) are used.[2]

3.2.3 Variable Selection for Random Forest using the Boruta Algorithm

For selection of explanatory variables to be used for Random Forest, the Boruta package available for usage in R is utilized. For each forecasting horizon, the full dataset containing all 299 variables considered is run on the Boruta algorithm. Since the ranking of only the top ranked variables are of interest, maximum number of iterations is restricted to 100 iterations. The mean of the Z-scores from the 100 iterations are extracted from the results of the Boruta algorithm and the variables are ranked based on this measure.

The number of variables confirmed to be 'important' by the Boruta algorithm are 74, 72, 76, and 65 for the 1-day-ahead, 5-days-ahead, 10-days-ahead and 22-days ahead forecast settings, respectively. In this paper, the list of top-20 ranked variables are reported and utilized for the purpose of selection of the optimal set of variables to forecast VIX in a random forest setting.

[2] The 14 variables are S&P k-day return, S&P 500 volume change, oil k-day return, trade-weighted USD change, credit spread and term spread with k = 1,5,10,22,66.

4. RESULTS

4.1 Variable Ranking by the Boruta Algorithm

Table 2 lists the rankings of the variables determined from the Boruta algorithm for each forecasting horizon.

Overall, the variability in the variable rankings among the different forecasting horizon settings does not seem to be large, especially for the variables ranked among the top 10. For all forecasting horizon settings, the lagged value of logarithm of VIX recorded the largest mean Z-score with quite a margin from the exogenous variables. Also for all forecasting horizon settings, the top-2 ranked exogenous variables were credit spread and term spread (credit spread ranked first for 1/5/10-day(s)-ahead forecast setting and term spread ranked first for the 22-days-ahead setting). The list is followed by the continuously compounded multiple-days-returns on S&P 500 and oil futures and the daily change rate in the S&P 500 index and the volume of S&P 500.

The difference in rankings among the forecasting horizons seems to be more visible for variables ranked $\mathbf{11}^{th}-\mathbf{20}^{th}$, most of which are composed of daily returns on other stock market indices and daily returns on individual prices of S&P 500 composites. For the 1/5/10-day(s)-ahead forecast setting, the daily returns on the NASDAQ-100, Dow Jones Industrial Average and DAX Performance index made it into the top 20 list. Comparatively, the list for 22-

days-ahead forecast setting seems to be dominated by the daily return series of the individual S&P 500 composites.

These rankings are used for the selection of the best subset of variables. That is, the optimal number of the top ranked variables are considered as the dataset to be used for random forest.

Table 2: Variable Rankings Determined by Boruta

	1-day-ahead	5-days-ahead	10-days-ahead	22-days-ahead
0	CBOEVIX(PI)	CBOEVIX(PI)	CBOEVIX(PI)	CBOEVIX(PI)
1	Credit Spread	Credit Spread	Credit Spread	T10Y3M
2	T10Y3M	T10Y3M	T10Y3M	Credit Spread
3	SP_66day	SP_66day	SP_66day	SP_66day
4	SP_22day	SP_22day	Oil_66day	Oil_66day
5	SP_10day	Oil_66day	SP_22day	SP_22day
6	SP_5day	SP_10day	SP_10day	SP_10day
7	Oil_66day	SP_5day	SP_5day	SP_5day
8	S&PCOMP(PI)	Oil_22day	Oil_22day	Oil_22day
9	S&PCOMP(MV)	S&PCOMP(PI)	S&PCOMP(PI)	@FITB
10	Oil_22day	S&PCOMP(MV)	S&PCOMP(MV)	U:BAC
11	DJINDUS(PI)	U:BAC	U:BAC	U:AIG
12	U:GE	DJINDUS(PI)	DAXINDX(PI)	@HBAN
13	U:BAC	U:GE	DJINDUS(PI)	U:KEY
14	NASA100(PI)	DAXINDX(PI)	@FITB	S&PCOMP(PI)
15	DAXINDX(PI)	@FITB	U:AIG	S&PCOMP(MV)
16	@FITB	U:AIG	U:GE	Oil_10day
17	U:AIG	U:USB	@HBAN	U:RF
18	U:USB	NASA100(PI)	NASA100(PI)	U:GE
19	U:RF	U:RF	U:USB	U:COO
20	U:OMC	FTALLSH(PI)	U:RF	U:STT

4.2 The Optimal Number of Variables

A cross validation procedure was carried out to find out the optimal number of variables to be used in a random forest setting. The procedure is straight forward: to start with a dataset with no exogenous variables using only the lagged values of VIX as input variable into random forest. From there, add one exogenous variable at a time based on the ranking decided by the Boruta algorithm in Table 2. Then, record the forecast error from each dataset and find the number of variables that produces the smallest forecast error.

Figure 3: Number of Variables and Forecasting Error

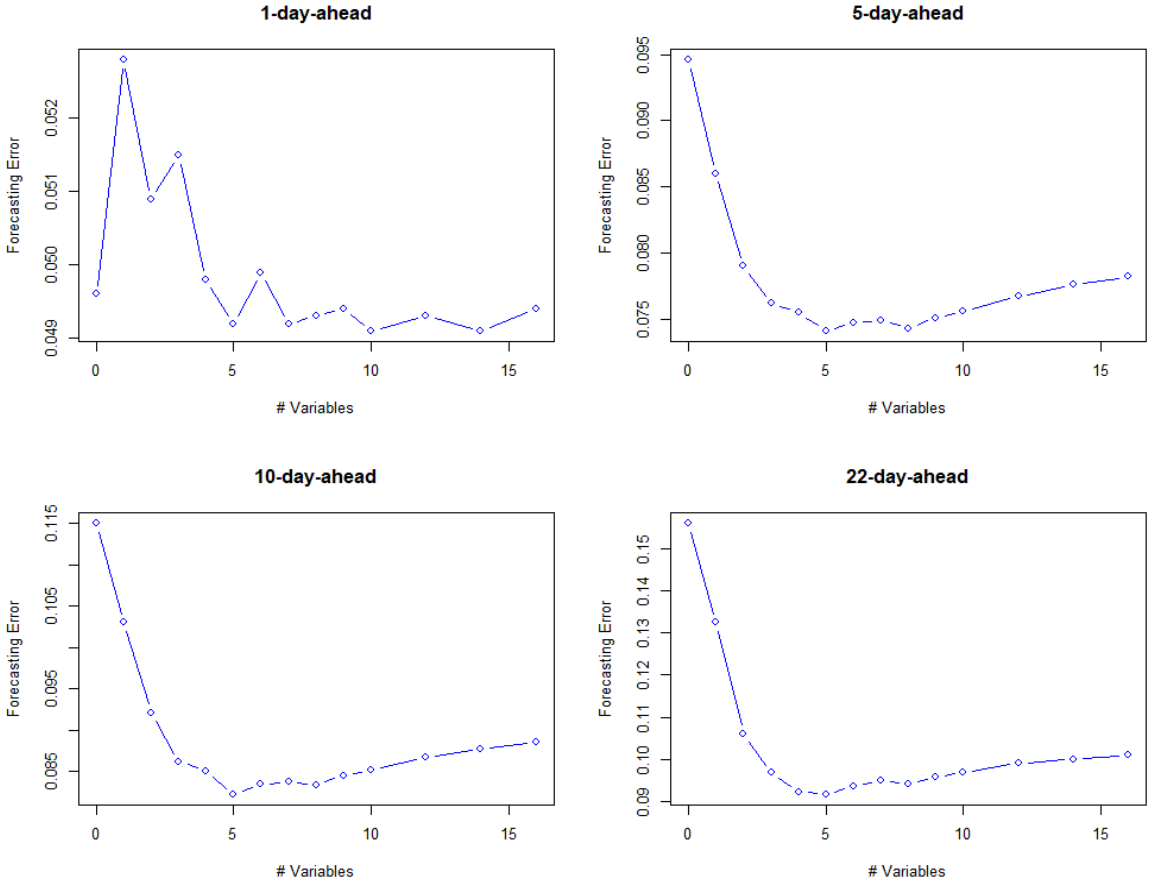


Figure 3 plots for each forecasting horizon setting the change in forecasting error as the number of variables included in the dataset increases. For the 1-day-ahead forecast, the picture seems less clear as it does not show a visible optimal point in terms of forecasting error. For the 5/10/22-days-ahead forecast setting, however, the results show a single minimal point along the plot of the forecast error. For the longer three forecasting horizon settings, using the 5 top-ranked exogenous variables result in the lowest forecast error.

Thus, among the 298 covariates that are considered, the best subset of variables is derived for forecasting VIX in a random forest setting; using the top-5 ranked variables according to the mean Z-scores derived from the Boruta algorithm.[3]

4.3 Forecasting Results

Table 3 shows the results of the out-of-sample forecasts of each model for the 1/5/10/22-day(s)-ahead horizons. The MSE, MAE and their relative ratios compared to the RW model are reported.

Among the five benchmark models from Fernandes et al. (2014), the pure HAR model shows the best performance overall, except for the 22-days-ahead forecast horizon where the NNHARX model records a smaller forecast error by

[3] In fact, for the 1-day-ahead setting the dataset with 10 variables produced the lowest forecast error. However, the difference to the dataset with 5 variables was negligible so only 5 variables were used to match with the other forecasting horizon settings.

a slight margin. Fernandes et al. (2014) claims that the relative success of the pure HAR model is due to the very persistent nature of the VIX index, and that it is pretty hard to beat the pure HAR process. Also, it can be said that the results show little evidence of non-linearity as the HAR model augmented with the neural network performs as well as the linear HAR model with no neural network (Ballestra et al., 2019).

Comparing the above results to the performance of random forest on different datasets show somewhat a different picture. Firstly, the RF(14) shows the results from random forest with dataset using only the exogenous variables used by the benchmark models in Fernandes et al. (2014). While the forecasting performance of RF(14) was even worse than the RW model for 1-day-ahead forecast, the forecasting error drops significantly for 5/10/22-days-ahead forecasts compared to the pure and augmented HAR models. The Diebold-Mariano-West (DMW) test statistic for comparison between HAR and RF(14) models for the 5-days-ahead forecasts show that the superior predictability of RF(14) to HAR is statistically significant at the 0.1% level. Considering that the gap between the two models is larger for 10/22-days-ahead forecasts, it can be said that RF(14) model is superior to the pure HAR model for the longer three forecasting horizons.

Moreover, the relative accuracy of random forest compared to the RW and linear benchmark becomes more evident as the forecasting horizon increases.

Table 3: Forecasting Performance at Different Horizons

	MSE	%	MAE	%	MSE	%	MAE	%	MSE	%	MAE	%	MSE	%	MAE	%
	One Day Ahead				*Five Days Ahead*				*Ten Days Ahead*				*Twenty-two Days Ahead*			
RW	0.0039	1.00	0.0456	1.00	0.0141	1.00	0.0891	1.00	0.0212	1.00	0.1105	1.00	0.0429	1.00	0.1544	1.00
ARX	0.0039	1.00	0.0447	0.98	0.0135	0.96	0.0876	0.98	0.0212	1.00	0.1106	1.00	0.0406	0.95	0.1500	0.97
HAR	0.0038	0.97	0.0445	0.98	0.0133	0.94	0.0873	0.98	0.0208	0.98	0.1098	0.99	0.0401	0.93	0.1502	0.97
HARX	0.0039	1.00	0.0446	0.98	0.0134	0.95	0.0871	0.98	0.0211	1.00	0.1101	1.00	0.0406	0.95	0.1488	0.96
NNHARX	0.0038	0.97	0.0446	0.98	0.0135	0.96	0.0874	0.98	0.0209	0.99	0.1088	0.98	0.0400	0.93	0.1484	0.96
RF(14)	0.0044	1.13	0.0491	1.08	0.0104	0.74	0.0767	0.86	0.0134	0.63	0.0858	0.78	0.0176	0.41	0.0989	0.64
RF(298)	0.0047	1.21	0.0507	1.11	0.0126	0.89	0.0854	0.96	0.0169	0.80	0.0989	0.90	0.0240	0.56	0.1180	0.76
RF(5)	0.0044	1.13	0.0492	1.08	0.0097	0.69	0.0741	0.83	0.0122	0.58	0.0822	0.74	0.0152	0.35	0.0917	0.59
RF(5*)	0.0044	1.13	0.0490	1.07	0.0118	0.84	0.0825	0.93	0.0170	0.80	0.0977	0.88	0.0272	0.63	0.1230	0.80

Notes: The forecasting performance of different models for the test period from February 29, 2000 to January 15, 2013 (3,240 daily observations). The results of the benchmark models (RW/ARX/HAR/HARX/NNHARX) are as reported in Fernandes et al. (2014).

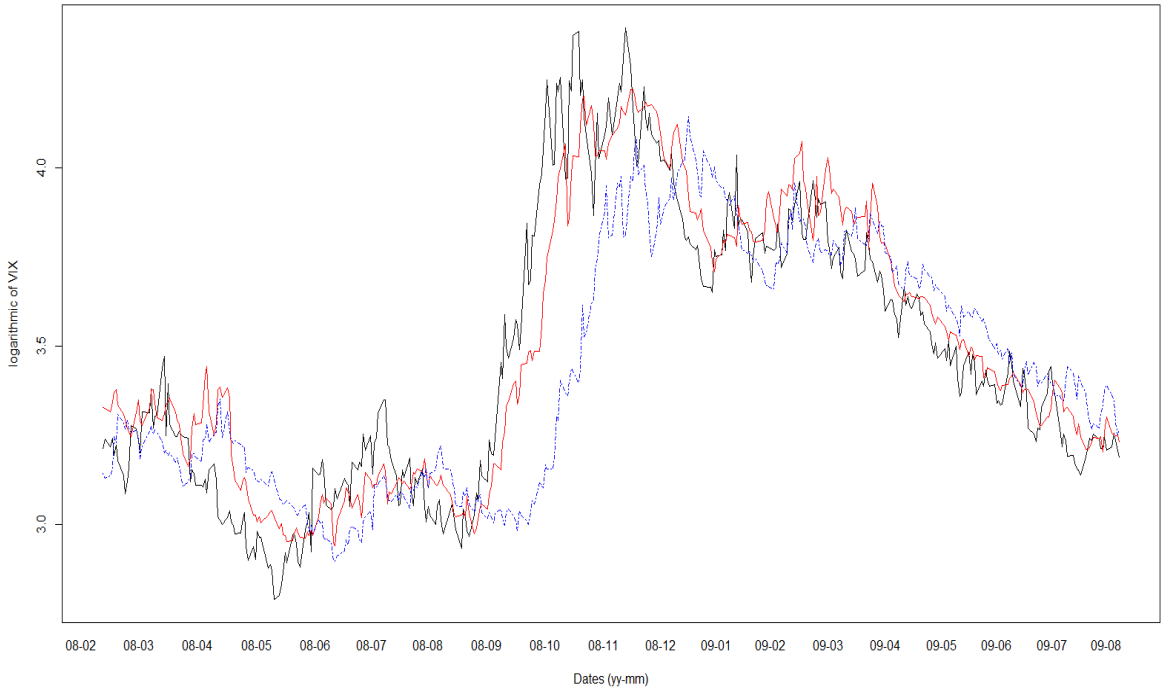
That is, the difference in forecasting error between the benchmark models and the RF models is enlarged as the forecasting horizon approaches the 30-calendar-days-ahead threshold. For RF(14), the relative MAE compared to RW model is 1.08, 0.86, 0.78 and 0.64 for the 1/5/10/22-day(s)-ahead forecasting horizons, respectively. This corroborates the results of Medeiros et al. (2019) which finds that the forecasting superiority of random forest compared to linear models becomes more evident for longer forecasting horizons.

RF(298) shows the performance of random forest when all 298 exogenous variables are included in the dataset, whereas RF(5) is the performance of random forest using the dataset of the optimal number of the most important variables, derived through the variable selection process described in sections 4.1 and 4.2. The results show that selecting the optimal number of the most important variables through the variable selection process using the Boruta algorithm further enhances the performance significantly. For the longer three forecasting horizons, RF(5) is able to produces results that are more accurate than RF(14). The DMW test statistic comparing the forecasts of RF(14) and RF(5) for the 5-days-ahead horizon show that the improvement in predictability is statistically significant at the 0.1% level. Since the difference in forecasting error between the two models is larger for the 10/22-days-ahead forecasts, it can be said that RF(5) is significantly more accurate than RF(14) for the longer three forecasting horizons.

RF(5*) model eliminates the two exogenous variables that were consistently found to be the most important for all forecasting horizon: credit spread and term spread. Thus, the model uses the five top-ranked variables after eliminating the top two from the list. It can be seen that the results of RF(5*) are deteriorated by a significant margin compared to the RF(5) model. Especially for the 22-days-ahead forecasting horizon, the relative MAE compared to the RW model increases from 0.59 to 0.80 when the top two variables are eliminated. This shows that only a small number of the most important variables can have an immense impact in the forecasting performance. Thus, it can be emphasized the importance and benefits of careful selection of covariates when forecasting using the random forest.

Figure 4 compares the forecasts of RF(5) and the HAR model for the period around the 2008 global financial crisis, from February 2008 to August 2009. The black line shows the actual value of logarithmic of VIX, which reached its all-time peak in October 2008. The red and blue lines show the 22-days-ahead forecast of RF(5) and HAR models, respectively. It can be seen that the forecast by RF are more accurate than that of the HAR model. RF catches much faster the sharp upward trend of VIX after the collapse of Lehman Brothers in September 2008, as well as the downward trend after the peak of the financial crisis. Overall, while the gap in forecasting error between the two models seem to be consistent over the whole forecasting period, it is in such highly volatile periods when the gap between the two models become much more evident.

Figure 4: Comparison of Forecasts Between RF and HAR Model



Notes: The logarithm of VIX from February 13, 2008 to August 12, 2009 (black line), along with the 22-days-ahead forecasts of RF(5) (red line) and HAR(blue line) models.

4.4 Robustness Check

As a robustness check, the same forecasts are made for the most recent period (November 28, 2018 to November 27, 2020, 504 daily observations) using a rolling window of the same size (2,500 daily observations). The benchmark models are the RW, ARX, HAR and HARX models. For random forest, the results from model using the five top-ranked exogenous variables from section 4.1 are reported (RF5).

The results are shown in Table 4. The results show quite the similar picture

Table 4: Forecasting Performance for 2018−2020 Period

	MSE	%	MAE	%	MSE	%	MAE	%	MSE	%	MAE	%	MSE	%	MAE	%
	One Day Ahead				*Five Days Ahead*				*Ten Days Ahead*				*Twenty−two Days Ahead*			
RW	0.0835	1.00	0.0593	1.00	0.1740	1.00	0.1237	1.00	0.2430	1.00	0.1748	1.00	0.3577	1.00	0.2505	1.00
ARX	0.0833	1.00	0.0585	0.99	0.1736	1.00	0.1233	1.00	0.2370	0.98	0.1688	0.97	0.3317	0.93	0.2302	0.92
HAR	0.0840	1.01	0.0580	0.98	0.1770	1.02	0.1217	0.98	0.2425	1.00	0.1622	0.93	0.3408	0.95	0.2210	0.88
HARX	0.0838	1.00	0.0584	0.98	0.1759	1.01	0.1247	1.01	0.2406	0.99	0.1692	0.97	0.3368	0.94	0.2298	0.92
RF(5)	0.0944	1.13	0.0641	1.10	0.1597	0.92	0.1119	0.90	0.1913	0.79	0.1274	0.73	0.2356	0.66	0.1514	0.60

Notes: The forecasting performance of different models for the test period from November 28, 2018 to November 27, 2020 (504 daily observations), using a rolling window of 2,500 daily observations. The results of the benchmark models (RW/ARX/HAR/HARX) are derived using the same model specifications as reported in Fernandes et al. (2014).

with the main results of the paper in section 4.3. Among the benchmark models, the pure HAR model records the lowest forecasting error in terms of MAE for all forecasting horizons. Its relative MAE compared to the RW model is 0.98, 0.98, 0.93 and 0.88 for the 1-day, 5-days, 10-days and 22-days-aheadforecasts, respectively.

As for RF(5), it records a higher forecasting error than the RW model for the 1-day-ahead forecasting horizon. For the longer forecasting horizons, however, its accuracy compared to the RW and the linear models is superior. Moreover, the gap between RF and the benchmark models are widened as the forecasting horizon becomes longer. The relative MAE of RF(5) compared to the RW model is 0.90, 0.73 and 0.60 for the 5-day, 10-day and 22-day-ahead forecasts, respectively.

Thus, the relative performance among the models on forecasting the most recent period seems to be in line with the main findings of the paper in the previous sections.

5. CONCLUSION

This paper seeks to apply random forest and its variable importance measure to forecasting the CBOE Volatility Index (VIX). Especially, it seeks to improve the multi-days-ahead forecasting of VIX compared to those reported in the

previous literature.

Compared to results of Fernandes et al.(2014) which finds it is very hard to beat the pure HAR process in forecasting VIX, random forest could produce forecasts that are significantly more accurate than the HAR and augmented HAR models for multi-days forecasting horizons. Moreover, the superior predictability of random forest compared to the RW and benchmark linear models becomes more apparent as the forecasting horizon becomes longer. This is in line with Medeiros et al.'s (2019) findings in the context of forecasting US inflation.

Further improvements in forecasting performance is attained through a systematic selection of covariates among a dataset consisting of 298 exogenous variables. Utilizing the Boruta algorithm, the rankings of the variables are extracted based on the permutation importance measure of random forest. Selecting the optimal number of the most important variables enhances the forecasting accuracy of random forest significantly. It seems clear that variable selection functions as a crucial factor for the predictability of random forest.

The robustness of the main results of the paper are confirmed through forecasting on the most recent period from 2018 to 2020.

While this paper focuses solely on the random forest, it would be interesting to investigate other ML methods that can capture nonlinear characteristics of VIX, especially deep learning methods such as the long short-term memory.

References

Ahoniemi, K. (2006). Modeling and forecasting implied volatility: An econometric analysis of the VIX index. *Working paper*, *Helsinki School of Economics.*

Ballestra, L. V., Guizzardi, A. & Palladini, F. (2019). Forecasting and trading on the VIX futures market: A neural network approach base on open to close returns and coincident indicators. *International Journal of Forecasting*, 35, 1250-1262.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone, (1984). *Classification and regression trees,* Wadsworth Books.

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

Degiannakis, S. A. (2008). Forecasting VIX. *Journal of Money, Investment and Banking*, 4, 5-19.

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

Hastie, T., Tibshirami, R. & Friedman, J. (2001). *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer.

Kohavi R, John GH (1997). Wrappers for Feature Subset Selection. *Artificial Intelligence*, 97, 273-324.

Konstantinidi, E., Skiadopoulos, G., & Tzagkaraki, E. (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.

Kursa, M. B. , & Rudnicki, W. R. (2010). Feature selection with the Boruta package, *Journal of Statistical Software*, 36, 1-13.

Medeiros, M. C., Vasconcelos, G. F., Veiga, A., and Zilberman, E. (2019). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, 1-45.

Mullainathan, S. & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31, 87-106.

Psaradellis, I., & Sermpinis, G. (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.