FORECASTING GOVERNMENT BOND YIELD BY MACHINE LEARNING

Phan Huy Tâm[1], Nguyễn Đức Minh Tấn[2], Nguyễn Vũ Thiên[3], Hà Mỹ Duyên[4], Nguyễn Việt Thường[5], Huỳnh Thủy Tiên[6]

[1] First and Corresponding Author, Ph.D. Candidate, Faculty of Finance and Banking, University of Economics and Law, VietNam National University, Ho Chi Minh City, Vietnam.
Email: tamph@uel.edu.vn

[2] Corresponding Author, Bachelor Student, Faculty of Finance and Banking, University of Economics and Law, VietNam National University, Ho Chi Minh City, Vietnam.
Email: tanndm19414c@st.uel.edu.vn

[3] Bachelor Student, Faculty of Finance and Banking, University of Economics and Law, VietNam National University, Ho Chi Minh City, Vietnam.
Email: thiennv19414c@st.uel.edu.vn

[4] Bachelor Student, Faculty of Finance and Banking, University of Economics and Law, VietNam National University, Ho Chi Minh City, Vietnam.
Email: duyenhm19414c@st.uel.edu.vn

[5] Bachelor Student, Faculty of Finance and Banking, University of Economics and Law, VietNam National University, Ho Chi Minh City, Vietnam.
Email: thuongnv19414c@st.uel.edu.vn

[6] Bachelor Student, Faculty of Finance and Banking, University of Economics and Law, VietNam National University, Ho Chi Minh City, Vietnam.
Email: tienht19404c@st.uel.edu.vn

ABSTRACT: This research aims to predict government bond yields in Vietnam based on macroeconomic factors that are believed to affect the bond market. The study was conducted based on a dataset spanning from July 2006 to December 2019 and is based on two samples of 1-year and 5-year government bonds, respectively, for short-term and long-term debt instruments. By using traditional predictive models such as Multiple Linear Regression, Vector Autoregression, Autoregressive Integrated Moving Average and some Machine Learning models such as Decision Tree, Random Forest and Support Vector Regression. The research objective of the authors is to find the most effective model for forecasting government bond yields. The results of the study show that Machine Learning models still have the advantage in both forecasting performance and model execution time. Among the Machine Learning methods implemented, the SVR model is the model with the best predictive performance for short-term and long-term government bond yields.

KEYWORDS: Government bond yields, Machine Learning, Vietnam.

I. INTRODUCTION

The Government bond market is an important part of the financial market, playing the role of an important long-term capital mobilization channel for the state budget, linking the issuance of government bonds with restructuring government debt, through focusing on issuing long-term bonds and diversifying the investor base (TRINH, NGUYEN, NGUYEN, & NGO, 2020). Government bonds and Government-guaranteed bonds are "commodities" of this market, with low-risk characteristics, making an important contribution to creating a common ground for yields in the financial market in general. Besides, when the government bond market develops, it will contribute a lot to the Government in implementing policies to regulate the economy by macro policies (Valko, Marques, & Castellani, 2005).

Because of its low-risk and high-potential characteristics, the Government Bonds market possesses a risk-free term structure, creating a standard reference base for different economic entities in the financial market to Bond issue pricing of corporate bonds or other financial products (TRINH et al., 2020). In addition, according to the development trend of technology, Machine Learning models have the potential to become an effective assistant, powerful analytical tool, and promising in effective investment management (Gan, Wang, & Yang, 2020). At the same time, after reviewing typical studies on the topic of government bonds, the team found that there was a lack of studies using data from developing countries. And the reason is that data sources in developing countries are relatively new, thin and difficult to apply advanced techniques to research.

This study aims to predict government bond yields in Vietnam based on macroeconomic factors that are believed to affect the bond market using machine learning algorithm such as: Decision Tree, Random Forest and SVR, based on data sources from developing countries, specifically Vietnam, and at the same time, compared with the forecast results of linear regression models such as ARIMA and OLS, to find the most optimal solution in forecasting the volatility of government bond yields.

II. LITERATURE REVIEW

Forecasting government bond yields has always been the interest of researchers for many decades. Therefore, there have been many classic theories of yield curve simulation. Typically, Ludvigson and Ng (2009) use the results of Nelson and Siegel's research to simulate the US bond curve. Reschreiter (2003) also applied APT in UK government bond yields and found that economic and financial risk factors are related such as inflation, changes in volatility slopes of the term structure, retail sales growth, and profitable stock market performance.

In the history of research on government bond yields, the studies of Ludvigson and Ng (2009), Cooper and Priestley (2009), and Cieslak and Povala (2015) demonstrate the predictive power of macro variables on bond surplus. Besides, Ghysels, Horan, and Moench (2018) show that the ability to predict macro variables will be limited, while Campbell and Shiller (1991) argue that the current yield differential has predicted future excess returns and profits. When establishing a model for forecasting government bond yields, macroeconomic factors also play an important part in building a robust model. If macro variables are not included in the model, the risk premium is only cyclical in nature, but according to the actual estimation results, the risk premium is clearly countercyclical. According to Ludvigson and Ng (2009), investors should be partially compensated for the risks caused by the macroeconomic downturn. In addition, Ludvigson and Ng (2009) also showed that the inclusion of macroeconomic components in the model will increase the forecast performance after combining 132 macro variables. Moreover, many researchs have also provided much evidence for this point such as Cooper and Priestley (2009) and Cieslak and Povala (2015).

However, in reality, in each country or in different economies, the macroeconomic variables that affect government bond yields are different. It is very difficult to apply the macroeconomic factors of an economic power like the United States or Europe to countries with thin economies like Southeast Asian countries. Therefore, it is necessary to find out the factors that effectively affect the yield and can strengthen the model to work more effectively.

In understanding the factors affecting government bond yields, a number of studies have shown that there is a great influence from factors such as exchange rate, budget deficit with a certain lag and interest rate. It can be followed by studies of (Akram & Das, 2019), (Kluza & Sławiński, 2002); (Cherif & Kamoun, 2007); (Agnihotri, 2015); (Yieand & Chen, 2019). In addition, in the studies of Ludvigson and Ng (2009); (Song, 2017); (Cherif & Kamon, 2007); Yieand & Chen, 2019 also applied variables such as GDP growth, industrial production index, inflation, monetary policies.

In previous studies by Ludvigson and Ng (2009); Cherif and Kamoun (2007); Yie and Chen (2019) also applied variables such as GDP growth to forecasting government bond yields. Strong economic growth tends to boost the performance of private investments and equity, while bonds are an alternative to capital-market investment that would have to provide higher yields to generate bonds to attract more investors. Moreover, economic growth always has a consequential effect on the government bond curve. Therefore, when there is a change in the country's GDP, there will be strong effects on government bonds in the future.

In addition, the above studies also mention the inflation variable, and the monetary policy. Inflation affects interest rates, when the inflation rate increases, if the real interest rate is to be stable and positive, the nominal interest rate must increase according to the inflation rate. The increase in nominal interest rates will lead to the consequences that the economy has to suffer, which is an economic recession and increased unemployment. Inflation is also a barrier for investors, affecting their long-term investment preferences, and bond yields tend to increase due to higher risk. Government fiscal variables affect long-run interest rates in standard Keynesian IS-LM models. Besides, it also shows the importance of short-term interest rates in promoting long-term interest rates.

In addition, the results from the study of Kumar and Baldacci (2010); Santosa & Sihombing, 2015; Zaja, Jakovcevic, and Visic (2018) and Inoguchi (2007) also show that the change of factors like exchange rate, inflation, foreign exchange reserves, US government bond interest rates and political variables also fluctuate government bond yields. According to research by Yieand and Chen (2019) on the Malaysian market, the author proves that foreign interest rates have a significant impact on explaining the volatility of government bond yields when bond yields reflect corresponding to the FED interest rate (USA) for a short period of time. However, the influence of this variable varies from country to country. The reason for this effect is that some investors may be sensitive and vulnerable from the change in US monetary policy while others are not. As a result, the situation could affect government bond yields immediately in the medium term, or without any reaction.

In studies on government bonds, risks are also mentioned by Gilles and LeRoy (1991) that have an impact on government bonds, including risks arising from macro factors and risks from major fluctuations caused by bond yields. According to Ul Haque, Mark, and Mathieson (1998) and Han, Kang, and Shin (2016), the authors of these researchs argue that the lower the credit ratings of the bonds, the higher the return. Furthermore, investors are willing to pay a lower risk premium for bonds with higher institutional ownership, but a higher yield for bonds with greater equity in the form of receipts depository.

There are a number of studies that have an interesting view on the predictability of government bond yields. The studies of Ludvigson and Ng (2009), Cooper and Priestley (2009), and Cieslak and Povala (2015) demonstrate the predictive power of macro variables on bond surplus. At the same time, Ghysels, Horan, and Moench (2018) show that the ability to predict macro variables will be limited. Campbell and Shiller (1991) argue that the current yield differential can predict future excess returns and profits. Ang and Piazzesi (2003) added the main components of groups of macro variables such as inflation and economic growth, along with latent variables to the VAR term structure model and found that these factors have an affect bond prices and can contribute to a significant improvement in yield prediction. In addition, there are a number of other studies that include macro variables in term structure models such as: Cooper and Priestley (2009), Bansal and Shaliastovich (2013), Greenwood and Vayanos (2014), and Cieslak and Povala (2015). Therefore, the authors find a significant influence of the term structure on the price as well as the change of government bond yields.

In addition to research papers on developed markets in the world, there are also many other studies on government bond yields in countries with developing economies. In Indonesia, Santosa and Sihombing (2015) produced a study on how volatility in interest rates and stock indexes has greatly influenced the slope of government bond yields through macro factors. This study used the Vector Error Correction Model (VECM) to overcome the weaknesses of the VAR model in analyzing the contribution of factors affecting the volatility of the yield curve, it has been shown that the curvature is mainly influenced by the volatility of interest rates, and finally factors such as exchange rate, inflation and foreign exchange reserves also contribute significantly to the volatility of government bond yields. In Vietnam, TRINH et al. (2020) has demonstrated the change of Vietnamese government bond yields based on the theory of term structure of interest rates. This study also shows that bond yields fluctuate due to macro variables such as basic interest rates, foreign interest rates, stock market returns, fiscal deficit, public debt, current account balances and prior period yields.

In addition to traditional studies or the application of simple models, breakthrough studies using new and modern tools also support the research process, especially in forecasting government bond yields. In the field of Machine Learning research, we can divide into two main methods that are supervised and unsupervised learning. Most research in practice tends to use supervised learning techniques. This technique will try to learn and build the best possible reflection function from the input data provided by the user and make a forecast of the output data. The popular methods of supervised learning methods such as Random Forest, Decision Tree, Neural Network. The unsupervised learning technique is that you just need to give the input data and do not have any output variables for the model. The core purpose of this technique is to model the data structure or distribution of data to better understand the nature of the model. Principal Component Analysis (PCA) is considered to be one of the classic examples of unsupervised learning, although it is considered to be non-Machine Learning (Hastie, 2009).

The use of Machine Learning contains disadvantages and advantages. However, the technique could help researchers perform calculations that are not possible for humans. Besides, Machine Learning can also increase the time and productivity of model operation. Because of the effectiveness of these Machine Learning methods in forecasting output data, there have been many research papers that have applied this method to forecast future government bond yields. In addition, Mullainathan and Spices (2017); Athey and Imbens (2019) based on the development of Machine Learning technical statistical properties and theoretical properties of Deep Neural Networks, many papers have applied them to their research in the field of economics. Suimon (2018) built a yield curve model based on Machine Learning and demonstrated the yield term structure based on the relationship of the three yield periods of 5, 10 and 20 years of the yield curve instead of the three-factor Nelson-Siegel three-factor models. He then used Short-Term Memory (LSTM) to predict long-term yields. Extending this study and analysis, Suimon et al. (2019a, 2019b) combined the yields of two countries, Japan and the United States, into the neural network model. In addition, based on the theory of interest rate parity, they combined the Dollar - Yen exchange rate along with US and Japanese interest rates into the above neural network model. Bianchi, Büchner, and Tamoni (2021) reviewed and evaluated a number of Machine Learning calculations to predict abundant returns on US Treasury bonds. This analysis has methods: Linear Regression, Partial Least Squares Regression (PLS), Decision Tree, Random Forest and Neural Network. By extension, Bianchi et al. (2021) found that the Neural Network method and the Decision Tree method performed the best in all the considered applications.

There are also studies demonstrating that non-linear methods can favorably predict bond excess returns. The Extreme Tree method and Neural Network were developed by Bianchi et al. (2021) applied to build a bond risk prediction model. Specifically, in the field of macroeconomics, Bianchi et al. (2021) and Huang and Shi (2011) use Machine Learning techniques to create models to evaluate the predictability of bond returns. Huang and Shi (2011) used LASSO method - a regression analysis method that performs variable selection and adjustment to improve the prediction accuracy and interpretability of the resulting statistical model and to monitor, capture macroeconomic risks, and from there it is possible to construct a single non-disparity factor from a system of 131 macro variables. Furthermore, Farrell, Liang, and Misra (2021) used the Deep Neural Network method and established valid inference about finite parameters following the first step using Deep Learning. In some recent articles by Gu, Kelly, and Xiu (2021); Chen, Pelger, and Zhu (2020); Feng, He, and Polson (2018); Feng, Polson, and Xu (2018); Heaton, Polson, and Witte (2016), they studied and applied Deep Learning and some variations of Deep Neural Networks.

III. DATA & METHODOLOGY

Data

The author's empirical study is based on a monthly dataset of Vietnamese Government bond prices with par value of VND and fixed interest rates, from July 2006 to December 2019 combined with Vietnam's macro variables. The authors study based on short-term and long-term debt instruments, respectively, with 1-year and 5-year terms of Government bonds. Based on previous empirical studies on government bond yields, especially in Vietnam, the authors build a model based on two dependent variables, one-year and five-year government bond yields, and other independent variables are the basic macroeconomic variables including Inflation Rate (CPI), Foreign Exchange Reserves (FER), Prime Interest Rate (Baserate), Foreign Interest Rate (FIR), Stock Return (stockreturn), total Government Budget Surplus or Deficit (Fiscal), Public Debt (Govdebt), Current Account Balance (CurentAcc). The data set used by the authors is based on the dataset of TRINH et al. (2020) studied the factors affecting the volatility of Vietnamese government bond yields.

Model Specifications

There are many previous studies that research on methods of forecasting government bond yields that apply many different regression analysis techniques such as OLS, VAR, ARIMA, etc. However, these traditional models still have shortcomings and must depend on the assumptions of each model or cannot be handled for non-linear models. The above defects can be solved by using Machine Learning models. In the study on forecasting government bond yields, the authors only use Supervised Machine Learning models. The authors divided into 2 settings to perform the forecast. Setup (1) uses representative models of traditional models such as ARIMA and OLS for forecasting. Set up (2) the authors forecast government bond yields by using Machine Learning models such as Random Forest, Decision Tree, and SVR with data of 1-year and 5-year bond yields combined with basic macro variables.

Autoregressive Integrated Moving Average (ARIMA)

The process of ARIMA model is understood as a model that combines three main factors: the autoregressive component AR (p), the stationarity of the time series I (d) and moving average model being the full name of MA (p). However, in this study, we use the ARIMA method to find out the parameters "p", "d", "q" automatically based on finding the ARIMA model (p,d,q) has the lowest AIC (Akaike Information Criterion). The benefit of this method is to avoid data overfitting (Nguyen Anh Phong, 2020). Theoretically, the model represents a multivariable linear regression equation of the input variables which includes the following main components:

Autoregression (AR): The regression component includes the lags of the current variables. The p-order delay is the retrograde value p time step of the sequence. The delay will be long or short of AR depends on the parameter p.

$$\text{(AR): } AR(p) = \emptyset_0 + \emptyset_1 x_{t-1} + \emptyset_2 x_{t-2} + \ldots + \emptyset_p x_{t-p}$$

Moving Average (MA): This process is understood as the process of moving or changing the average value of the series over time. As mentioned above, a series is stationary, so the mean change process is a white noise series. The MA process works by finding a linear relationship between the random components ϵ_t which should satisfy:

- $E(\epsilon_t) = 0$ ensure that the series is stationary, not changing the mean over time.
- $\sigma(\epsilon_t) = 0$ The variance of the series is constant.
- $\rho(\epsilon_t, \epsilon_{t-s}) = 0$

The combination of MA, AR and the stopping factor of the value chain I, identify the ARIMA model (p,d,q) with the order of autoregression p, d being the order of differencing of the survey time series and the moving average q:

$$\Delta x_t = \emptyset_1 \Delta x_{t-1} + \emptyset_2 \Delta x_{t-2} + \ldots + \emptyset_p \Delta x_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}$$

With: Δx_t is the d - order differencing and ϵ_t are the white noise series

The vast majority of time series have a correlation between the past value and the present, the higher the degree of this correlation, the closer the value of the series is to the present value. In addition, time series datasets are often seasonal, cyclical, and non-stationary. However, ARIMA only works for stationary data series. To perform model training, the data series needs to be normalized to stationary series by taking the difference or logarithm and checking for stationarity through tests such as ADF or KPSS. In this study, the authors found that the data set guarantees stationarity at first difference, that is, the parameter “d” = 1 is shown in Figure 1 and Figure 2.

The order of differencing: $I(d) = \Delta^d(x_t) = \underbrace{\Delta(\Delta(\ldots\Delta(xt)))}$.

d times

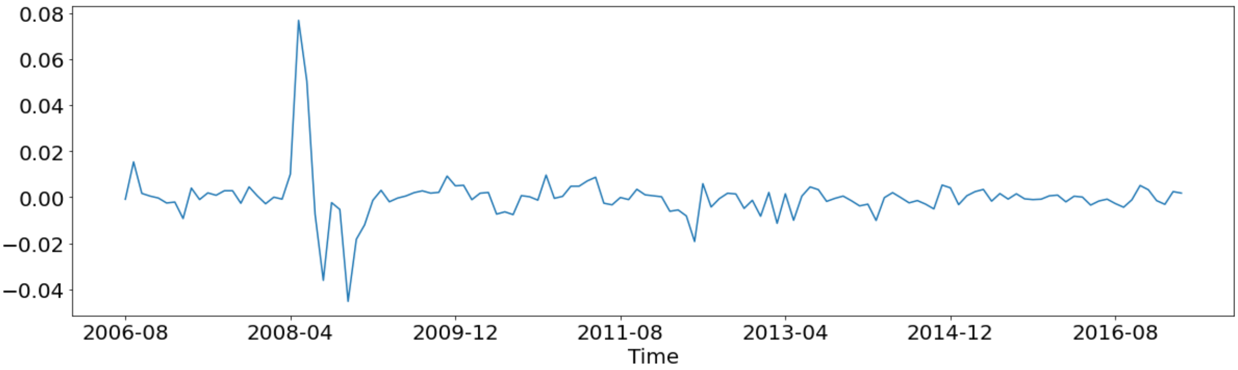

Figure 1. First order differencing of 1-year government bond yields

Source: Author's calculation

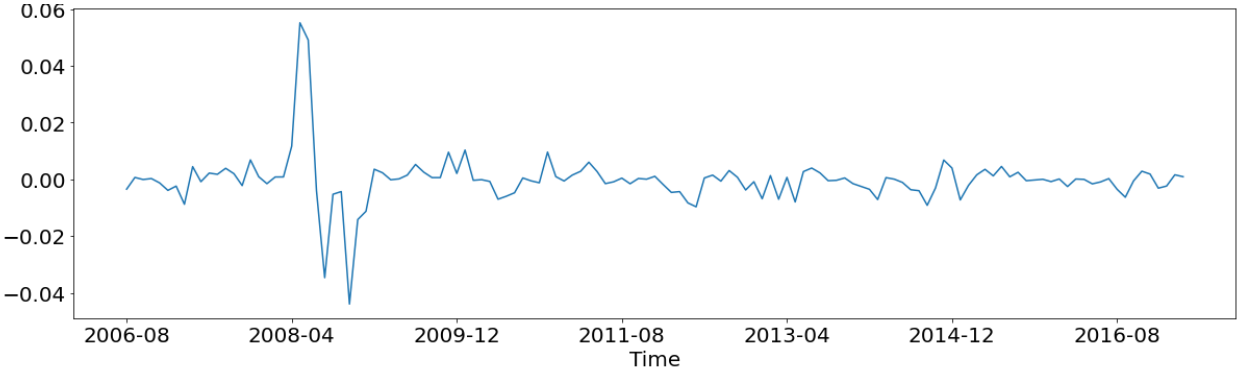

Figure 2. First order differencing of 5-year government bond yields

Source: Author's calculation

After determining the coefficient "d", we automatically find out the remaining two coefficients "p", "q" by testing the cases so that those two coefficients can produce model with the smallest AIC. In this study, the authors limit "p" and "q" to 5. After having the model, the authors use a new data set from May 2017 to December 2019 to evaluate the effectiveness that ARIMA brings in forecasting government bond yields. After the forecast results are available, the authors use error measurement indicators such as MAPE, RMSE. These two indicators will be used for all future models.

The above model is considered suitable for the research content because this model is applied on time series and the model is also based on the rule based on past values to predict the future. The main purpose of the model is to find a way to introduce lagged variables to create a predictive model that better fits the value of the series. However, this model needs to comply with one strict condition, that is, the error has no autocorrelation and the residual is white noise.

Multiple Linear Regression (OLS)

The authors used a classic Linear Regression model to examine the influence of macro variables on Vietnam government bond yields with 1-year and 5-year maturities with the following regression formula:

For short-term debt instruments: GB1Y = β0 + β1 GGVF5Y + β2 CPI + β3 FER + β4 Baserate + β5 FIR + β6 Stockreturn + β7 Fiscal + β8 Govdebt + β9 CurrentAcc

GGVF1Y = β_0 + β_1 GGVF5Y + β_2 CPI + β_3 FER + β_4 Baserate + β_5 FIR + β_6 Stockreturn + β_7 Fiscal + β_8 Govdebt + β_9 CurrentAcc

For long-term debt instruments: GGVF1Y = β_0 + β_1 GGVF5Y + β_2 CPI + β_3 FER + β_4 Baserate + β_5 FIR + β_6 Stockreturn + β_7 Fiscal + β_8 Govdebt + β_9 CurrentAcc

In there:
β_0: Constant.
β_1, β_2, …, β_9: The angle systems of the predictors of the respective macro variables.
GB5Y: Yield of 5-year government bonds.
GB1Y: Yield of 1-year government bonds.
CPI_t: is the inflation rate in month t.
FER_t: is non-gold monthly foreign exchange reserves (in USD) in month t.
$Baserate_t$: is the base rate of Vietnam in month t.
FIR_t: is the foreign interest rate (US interest rate) in month t.
$Stockreturn_t$: is the return of VN-Index in month t.
$Fiscal_t$: is overall budget surplus or deficit as a percentage of GDP in month t.
Govdebtt: is general government debt as a percentage of GDP in month t.
$CurrentAcc_t$: is current account balance as a percentage of GDP in month t.

Table 1. Expected impact of macro variables on government bond yields

Variables	Expectation
CPI	+
FER	-
Baserate	+
FIR	+
Stockreturn	-
Fiscal	+
Govdebt	+
CurrentAcc	+

Source: Synthesized from previous studies

Decision Tree

First and foremost, we must comprehend the notion of a decision tree, as well as the formula and benefits, in order to explain why we chose this model. The model is made up of subsets, each of which is a branch that is fully independent of the others and represents the value of an attribute to be examined. A decision node is produced when a subset splits into smaller subsets. The root node is the node that represents the complete dataset. Decision trees have the advantage of being able to handle both numerical and categorical data. In classification and regression problems, decision trees are frequently utilized. Instead of predicting a class in each node as in the classification problem, the regression method predicts a value.

The core of this method is to use a probability algorithm to separate the cases according to the set conditions and minimize the error. In the Regressive Decision Tree, the CART algorithm is used so that it tries to split the training set in a way that minimizes MSE instead of trying to split the training set in a way that minimizes non-significant data points.

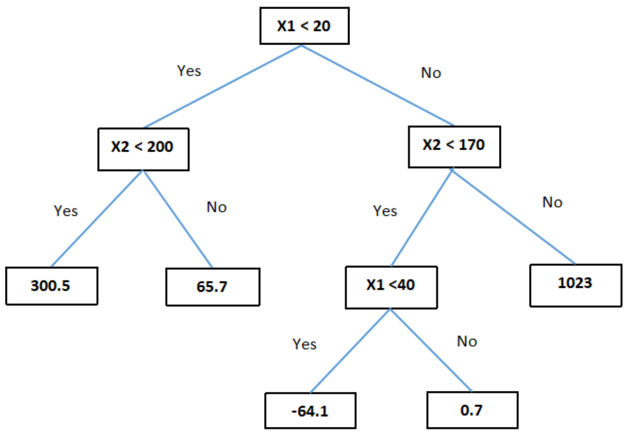

Figure 3. How the Decision Tree method works

Source: Author's synthesis

The cost function that the algorithm tries to minimize:

$$J(k, t_k) = \frac{m_{left}}{m} MSE_{left} + \frac{m_{right}}{m} MSE_{right}$$

$$In\ there: \begin{cases} MSE_{node} = \sum_{i \in node} (\hat{y}_{node} - y^{(i)})^2 \\ \hat{y}_{node} = \dfrac{1}{m_{node}} \sum_{i \in node} y^{(i)} \end{cases}$$

To perform model training and evaluation, the authors divided the original dataset into two parts. One part consists of 80% of the observations of the dataset to train the model, the other 20% to evaluate the effectiveness of the model on a certain random particle. The following Machine Learning methods also apply this setting.

For the following reasons, Decision Tree is the best way for making decisions. First, the problems are stated clearly, which aids in the prediction of the target variables. Second, the model's conclusions are thoroughly examined, as are the potential ramifications of the actions. Third, devise an effective mechanism for quantifying variables so that their outcomes and probabilities can be determined. Finally, the model is capable of working with enormous data fields and making the most accurate conclusions and predictions based on existing data. In conclusion, Decision Tree is a famous method that is extensively used due to its benefits and is ideal for predicting Vietnam government bond yields due to its interpretability and simplicity. However, the model is overly reliant on data and is prone to overfitting.

Random Forest

Random Forest is a supervised learning algorithm that works by building multiple decision trees based on different samples and making decisions according to the majority or the average of the classes.

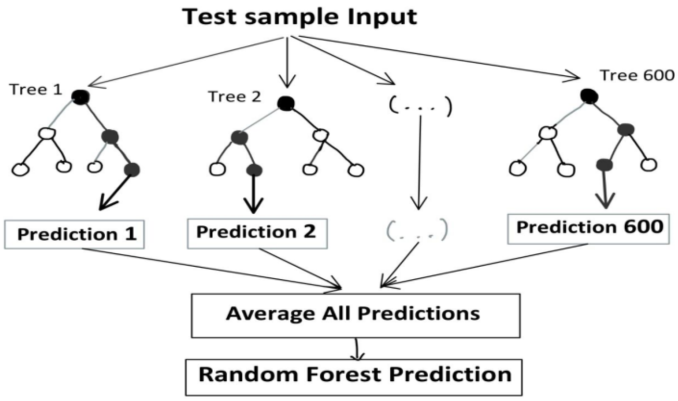

Figure 4. How the Random Forest method works

Source: Author's synthesis

Basic operation steps of Random Forest:

- Step 1: In the Random Forest, n random records are taken from a data set with k number of records
- Step 2: Individual decision trees are built for each sample
- Step 3: Each decision tree will produce an output
- Step 4: Final results are considered based on Majority or Mean Voting for classification and regression respectively

According to Rodriguez-Galiano, Sanchez-Castillo, Chica-Olmo, and Chica-Rivas (2015), Random Forest receives an input vector (x), which is made up of the values of the different obvious features analyzed. For a given training area, Random Forest builds a number of K regression trees and averages the results. After K trees $[T(x)]_I^K$ are planted, the Random Forest regression forecast is:

$$\hat{f}_{rf}^{K}(x) = \frac{1}{K} + \sum_{k=1}^{K} T(x)$$

To avoid correlation between trees, Random Forest increases the diversity of trees by growing them from different training data subsets generated through a process called Bagging. Bagging is a technique used to generate training data by resampling a random sample from the original data set with substitution, i.e. without removing the selected data from the input sample to generate the next subset {h (x, θ_k), k = 1,…, K}, where {θ_k} are independent random Vectors with the same distribution. As a result, some data may be used multiple times during training, while others may never be used. Because the prediction result of the random forest model is a combination of the results of the decision trees in the forest with low correlation, so the random forest has low bias and variance, achieving high stability. Another interesting feature is that the trees of the Random Forest classifier grow without branching, which makes them more lightweight from a computational perspective.

In addition, the samples not selected to train the Kth tree in the bagging process are included as part of another subset called out-of-bag (oob). These oob factors can be used to evaluate performance (Peters et al., 2007). In this way, Random Forest can compute an unbiased estimate of the generalization error without using a subset of the external text data (Breiman, 2001).

Support Vector Regression

Support Vector Machines (SVM) is one of the most popular and widely used algorithms for dealing with classification problems in Machine Learning. However, the use of support vector models in regression is not well documented. This algorithm acknowledges the presence of non-linearity in the data and provides a proficient predictive model using supervised learning algorithms and these types of models are called Support Vector Regression (SVR). In the SVR method, the required line to fit the data is called the hyperplane. The goal of an SVM algorithm is to find a hyperplane in a space of size n that can unambiguously classify the data points. The data points on either side of the hyperplane closest to the hyperplane are called the Support Vector. These vectors affect the position and orientation of the hyperplane and thus help to solve the regression optimization problem.

Some hyperparameters are used in SVR such as:

- Hyperplane: These are decision boundaries used to predict continuous output. The support vectors on either side of the hyperplane are used to draw the necessary line that shows the algorithm's prediction results.
- Kernel: A set of mathematical functions to find a hyperplane in a higher dimensional space to solve the problem of not being linearly separable for a data set containing two types of observations. a two-dimensional space. The most widely used kernels include functions such as Linear, Non-Linear, Polynomial, Radial Basis Function, and Sigmoid. Usually each of these kernels is used depending on the dataset, and often the default kernel used is RBF.
- Boundary Lines: These are two lines drawn around the hyperplane at a distance calculated by the factor ε (Epsilon). It is used to create a boundary between data points.

To understand how the modeling process works, it is first necessary to understand the underlying theory of the model briefly. With a training data set $\{(x_l, y_l), \ldots, (x_n, y_n)\} \subset \chi\, x\, R$, there are n number of observations included in the model training, where the input data domain is defined. In contrast to OLS, the objective function of SVR is to minimize the coefficients - more specifically, the L2 norm of the coefficient vector (L2-norm) - not the squared error. Instead, the error term is handled in the bound intervals known as the maximum error - ε. That is, the deviation on the y(i) of the whole training data set is not greater than ε. For the nonlinear regression case, the function fx can be defined as follows:

$$f(x) = \sum_{i=1}^{n} (a_i - a_i^*)\, K(x_i, x) \ + b$$

So that:

$$f(x) = \sum_{i=1}^{n} (a_i - a_i^*) = \ 0, \ \ v\grave{a}\ C \geq \ a_i, a_i^* \ \geq \ 0, \forall i$$

In that:

C: is a constant that determines the balance of error margin between the flatness of f and the accepted amount of excess deviation ε.

a_i, a_i^*: are Lagrange multipliers.

$K(x_i, x)$: is a Kernel function defined by the formula:

$K(x_i, x) \ = \ [\Phi(x_i), \Phi(x_j)]$ with Φ is the attribute mapping for the kernel K.

X_i are the input points with $[(a_i \ - a_i^*) \ \neq \ 0]$ are called the SVs.

Distinct kernel functions are utilized to generate different support vector regression for each different data characteristic, according to Wang and Xu (2017). RBF kernel functions are always extensively applicable, regardless of the conditions: small sample size, large sample size, or low sample size. The kernel function not only addresses the "curse" of low-dimensional Vectors being transferred to higher-dimensionality, but it also solves the size problem and enhances the machine's nonlinearity. Based on distinct kernel functions of SVR with variable generalizability, the kernel function has its own properties. Kernel functions are currently split into two categories: global kernel functions and local kernel functions. The local kernel function is useful for extracting pattern local characters. The value of the kernel function is influenced by the proximity of data points and the interpolation capabilities. As a result, learning ability is excellent. A local kernel function is the Kernel Gauss (RBF). The global kernel function is useful for extracting sample global characteristics. The kernel function is strongly generalizable because it is simply affected by the distance between the data points of the value. The global kernel function has a lower learning ability than the local kernel function. Global kernel functions include ar, Polynomial, and sigmoid. In summary, the local kernel function has good learning ability but poor generalization. For global kernel functions, the situation is reversed.

IV. RESULTS AND DISCUSSION

Descriptive statistics

The year 2006 was the beginning of the developing period of the stock market, the Ministry of Finance issued Decision 2276/QD-BTC on June 20, 2006 with the content that Hanoi Stock Exchange was used to bid for Government bonds, since then, creating a new turning point for Vietnam's government bond market. The first large-scale issuance of Government bonds had a total par value of up to one trillion Vietnam dong in early 2007. In the second issuance of 2007, the total par value doubled, and the winning rate was up to 78.5%. Each batch is issued in 3 batches. However, in the first stage, there were still many difficulties, the first failed bidding session quickly appeared in the second quarter of 2007, gradually reducing the winning rate in the last months of the year. Typically, in the bidding session on September 5, 2007, the winning rate was only 14.2%. The difficult situation of the market continued until the end of the second quarter of 2008. However, this period also witnessed a rare positive point, when there were sessions with the participation of foreign investors. Typically, in some sessions in March 2008, investors bought up to ten million Government bonds, accounting for nearly 90% of the volume of Government bonds traded on the market.

The bidding was not successful at this stage, leading to a series of unissued government bond issuance sessions. The reason is considered to be due to the low ceiling interest rate for a long time, although, until the beginning of 2008 the inflation situation increased, and deposit interest rates of banks also increased dramatically. However, this situation quickly ended from the second quarter of 2008 onwards, instead, the market gradually regained its attraction to investors. The 2008 financial crisis was an important event, influencing the Vietnamese government to rapidly develop into the domestic bond market in general and Government bonds in particular as a source of capital mobilization for projects and public investment, offset temporary state budget expenditure, and loans to businesses, financial institutions, and local governments.

In September 2009, the Government bond market officially came into operation on the Hanoi Stock Exchange. This is also a strong development of the government bond market, it plays an important role in the development of safe investment tools as well as contributing to the growth of the national economy. In the period from September 2009 to September 2014, the Hanoi Stock Exchange mobilized 654,493 billion VND for the State budget. The listed value of Government bonds in 2009 was about 152 trillion, the development by the end of 2019 is about 997 trillion, an increase of more than 6 times compared to 10 years ago.

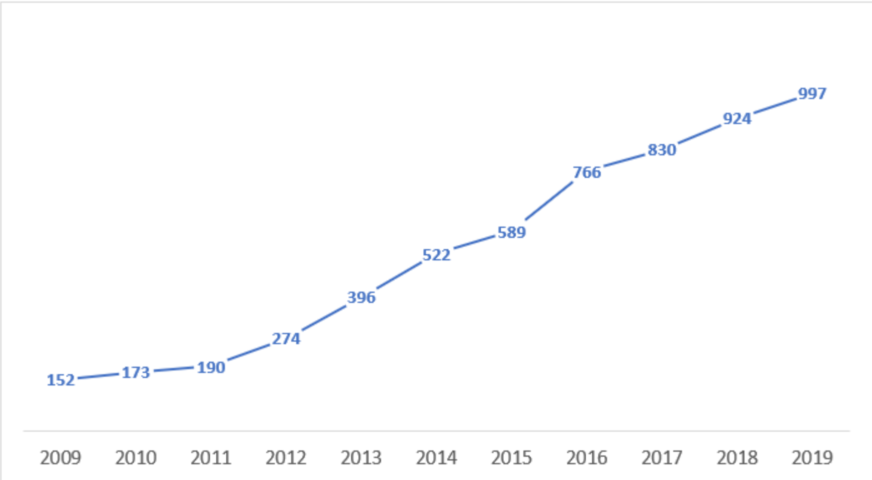

Figure 5. Listing scale of Vietnam Government bonds 2009-2019 (Unit: trillion VND)

Source: Hanoi Stock Exchange (HNX)

According to Trinh et al. (2020), the Vietnamese government set out a roadmap to increase the size of the bond market from about 20% in 2015 to 45% in 2020. However, Vietnam's deficit and public debt situation is increasing and is said to be less financially flexible. And by issuing more bonds, the government is expected to achieve the set roadmap, improving the weak points of the market. Besides, the volatility of Vietnam government bond yields from 2006 to 2019 is quite volatile. Specifically, the yield reached a peak of 20.33% in June 2008 and more than 12% in the first 3 months of 2010 and the last 6 months of 2011. While the yield was at the lowest level of about 2% at the end of 2019.

Table 2 shows the descriptive statistics of the data set that the group of authors collected 162 observations with the variables of 5-year government bond yield, 1-year government bond yield, inflation rate, foreign exchange reserves, base rate, foreign interest rates, stock returns, total government budget surplus or deficit, public debt, current account balance from July 2006 to December 2019.

Table 2: Statistical description of the variables

	Observations	**Mean**	**Standard deviation**	**Min**	**Max**
GB5Y	162	0.080716	0.033162	0.020000	0.203330
GB1Y	162	0.071499	0.036271	0.015500	0.211670
CPI (%)	162	0.601198	0.793960	-0.756000	3.912000
FER (billion $)	162	31.100000	1.630000	11.250000	68.810000
Baserate (%)	162	8.885802	1.085656	7.000000	14.000000
FIR (%)	162	1.479954	1.826785	0.310000	9.000000
Stockreturn (%)	162	0.764756	8.797782	-24.0092	38.517100
Fiscal (%)	162	-2.879439	1.856116	-5.027260	1.224610
Govdebt (%)	162	50.436752	6.870214	38.406899	59.662135
CurrentAcc (%)	162	-0.392593	5.182574	-11.00000	6.000000
					Source: Author's calculation

Through Table 2, we can see that the average yield of government bonds for the 5-year term is 8.07%, higher than the 1-year term of 7.15% in the period 2006 to 2019. This makes perfect sense because bonds with higher maturities tend to offer higher yields to investors. This is also explained because long-term bonds are generally riskier and more sensitive than short-term bonds. However, the sensitivity of government bond yields tends to reverse for the standard deviation index. In Table 1, the standard deviation of 5-year government bonds is 3.32% while the standard deviation of 1-year government bonds is the highest, equal to 3.62%.

Estimated results

The authors use Table 3 to report the correlation of macro-variable relationships with government bond yields and Table 4 to report the results of errors that the models used in the report are ARIMA, Multivariable Regression, Random Forest, Decision Tree, and Support Vector Regression.

Table 3: Correlation between variables

Government bond yield 1 year

	GB5Y	GB1Y	CPI	FER	Baserate	FIR	Stockreturn	Fiscal	Govdebt	CurrentAcc
GB5Y	1									

GB1Y	0.99	1								
CPI	0.4	0.41	1							
FER	-0.77	-0.71	-0.28	1						
Baserate	0.32	0.36	0.09	0.16	1					
FIR	0	0.05	0.21	0.04	0.12	1				
Stockreturn	-0.07	-0.08	-0.15	0	-0.16	0.02	1			
Fiscal	0.55	0.6	0.43	-0.47	0.24	0.59	-0.13	1		
Govdebt	-0.71	-0.72	-0.42	0.7	-0.02	-0.51	0.06	-0.8	1	
CurrentAcc	-0.44	-0.48	-0.36	0.41	-0.05	-0.52	0.11	-0.62	0.7	1

Government bond yield 5 year

	GB1Y	GB5Y	CPI	FER	Baserate	FIR	Stockreturn	Fiscal	Govdebt	CurrentAcc
GB1Y	1									
GB5Y	0.99	1								
CPI	0.41	0.4	1							
FER	-0.71	-0.77	-0.28	1						
Baserate	0.36	0.32	0.09	0.16	1					
FIR	0.05	0	0.21	0.04	0.12	1				
Stockreturn	-0.08	-0.07	-0.15	0	-0.16	0.02	1			
Fiscal	0.6	0.55	0.43	-0.47	0.24	0.59	-0.13	1		
Govdebt	-0.72	-0.71	-0.42	0.7	-0.02	-0.51	0.06	-0.8	1	
CurrentAcc	-0.48	-0.44	-0.36	0.41	-0.05	-0.52	0.11	-0.62	0.7	1

Source: Author's calculation

Table 4. Accuracy of prediction

Models	Government bond yield 1 year.		Government bond yield 5 year.	
	RMSE	MAPE	RMSE	MAPE
ARIMA	0.010538	33.06%	0.010168	25.7%
Multi-variate regression	0.007031	21.4%.	0.003548	8.28%
Decision tree	0.010737	9.3%	0.002776	8.21%
Random forest	0.006858	7.12%	0.006185	4.34%
SVR – RBF	0.018908	10.08%	0.017017	9.0%
SVR – Linear	0.003933	4.85%	0.003818	4.52%
SVR – Polynomial	0.015193	16.4%	0.015247	14.12%

Source: Model results from Python

For 1-year government bonds, the ARIMA model, after predicting government bond yields from May 2017 to September 2019 based on time series data from July 2006 to April 2017 gives us The trend in bond yields shown in Figure 6. However, the error between the predicted value and the actual value is still relatively high. Specifically, for the index RMSE = 0.010538 and MAPE = 33.06%. Similar to the 1-year term, ARIMA still forecasts the trend for the 5-year term in Figure 7, but the error between the actual value and the forecast value has been minimized. Specifically, the error indexes are RMSE = 0.010168, and MAPE = 25.7%, respectively. Although the forecast error of the 5-year term is smaller than that of the 1-year term, the error of the ARIMA model is still relatively high and has not yet given the optimal forecast results. Besides, previous studies by Ludvigson and Ng (2009) and Cieslak and Povala (2015) also show that incorporating macro variables into the model will increase forecasting efficiency. For that reason, the authors will continue to test methods that combine macro variables or Machine Learning methods on this same data set to find a new model that can better predict.

Figure 6. Actual and predicted 1-year government bond yields by ARIMA

Source: Model results from Python

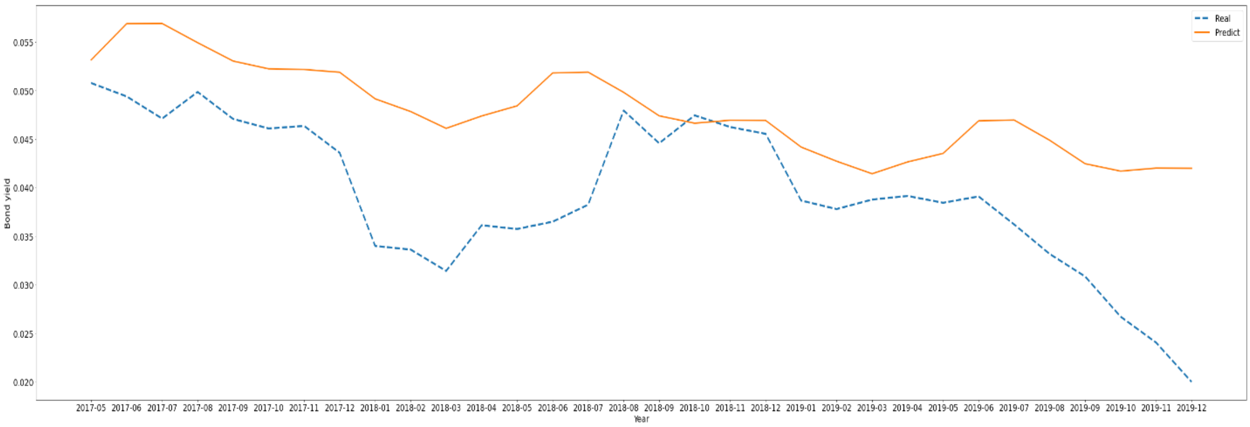

Figure 7. Actual and predicted 5-year government bond yields by ARIMA

Source: Model results from Python

After running the multi-variable regression model incorporating macroeconomic variables, the authors obtained the results showing the relationship of macro variables in Table 3. Through Table 3, we can see that Macro variables are divided into two groups: the first group includes variables that have a negative impact and the other group has a positive impact on government bond yields.

The short-term debt instruments are 1-year government bonds and the long-term instruments are 5-year government bonds, which are influenced mainly by the macro variables that the authors have included in the model. The variables such as other term bond yields, foreign exchange reserves, policy surplus or deficit, and public debt are the variables that have the strongest impact and are consistent with the expected shown in Table 1. However, two macro variables, US interest rates, and Vietnamese stock yields have no effect on government bond yields or have tiny effects. In addition, fiscal-related variables such as public debt and current account balances produce results contrary to the expectations in Table 1. This is contrary to the empirical results of the previous studies. Giordano, Linciano, and Soccorso (2012) and Kameda (2014) show that when public debt increases, a current account deficit can help boost government bond yields through a risk premium. However, there are still studies by NGUYEN (2019) and the TRINH et al. (2020) report that there may still be a relatively small negative relationship between the two variables for government bonds.

When combining macro variables and government bond yield value by multivariable linear regression model, we have a forecast result that is closer to reality than shown in Figure 8. Specifically, for term 1 year, index RMSE = 0.007031 and MAPE = 21.4%. This result has improved the forecast results of bond yields of the same term of the ARIMA model. The 5-year government bond yield also performed much better than the ARIMA model with RMSE = 0.003548 and MAPE = 8.28% visualized in Figure 9.

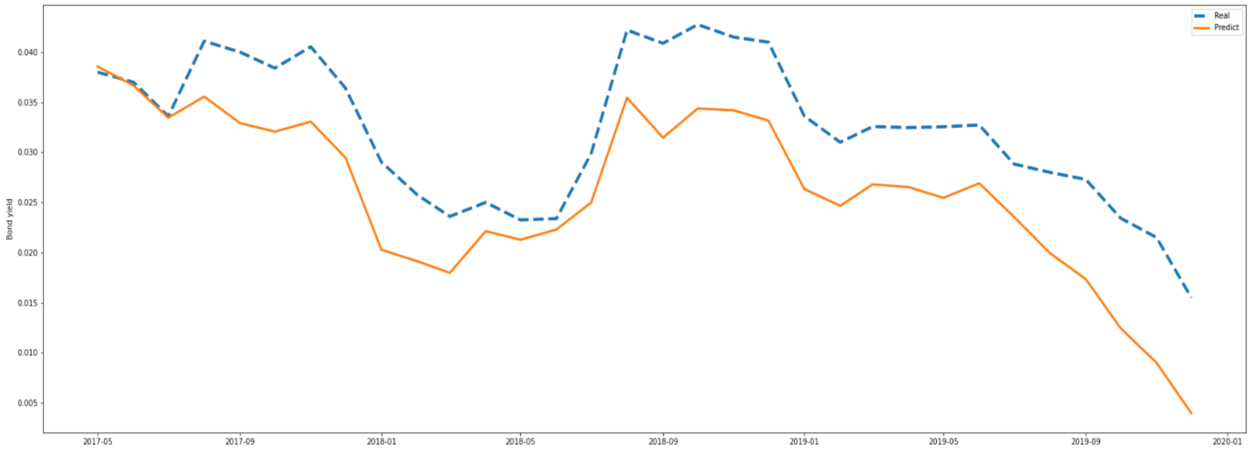

Figure 8. Actual and predicted 1-year government bond yields by OLS

Source: Model results from Python

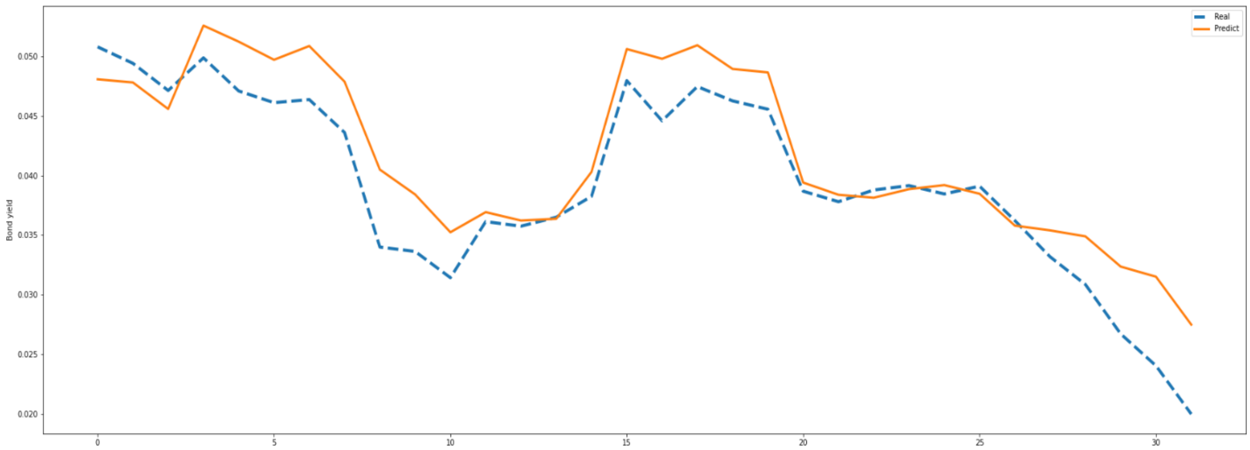

Figure 9. Actual and predicted 5-year government bond yields by OLS

Source: Model results from Python

Through the above analysis, it can be seen that it is necessary to add macro variables to the analysis and prediction of government bonds instead of just using historical data on bond yields. Although this is a traditional method that has been published for a long time, the regression method still gives acceptable results (Hoogteijling, 2020). In addition, this method not only improves the efficiency of the forecasting model, especially for long-term forecasts but also helps the reader of the research report to clearly see the relationship between the variables. Macro variables affect the yield of government bonds, from which it is possible to make judgments or make appropriate policies. However, this model still has certain limitations such as still having to comply with assumptions or not being able to optimize the residuals of the model well. In addition, Machine Learning models recently can bring the advantages of model computation time or high accuracy in forecasting. That is also the reason why the authors continue to experiment with predicting results with a number of Machine Learning models, and the authors expect that these Machine Learning models will bring more effective predictive results than other traditional models.

After applying the Machine Learning model to the prediction, the Decision Tree model gives better prediction results than the traditional regression model in the short and long term, which is clearly shown in Figure 10 and Figure 11. Specifically, based on Table 4, the Decision Tree model gives RMSE and MAPE for 1-year term at 0.010737 and 9.3%, respectively; the 5-year forward has an index of RMSE = 0.002776 and MAPE = 8.21%. This result is smaller than that of the traditional models.

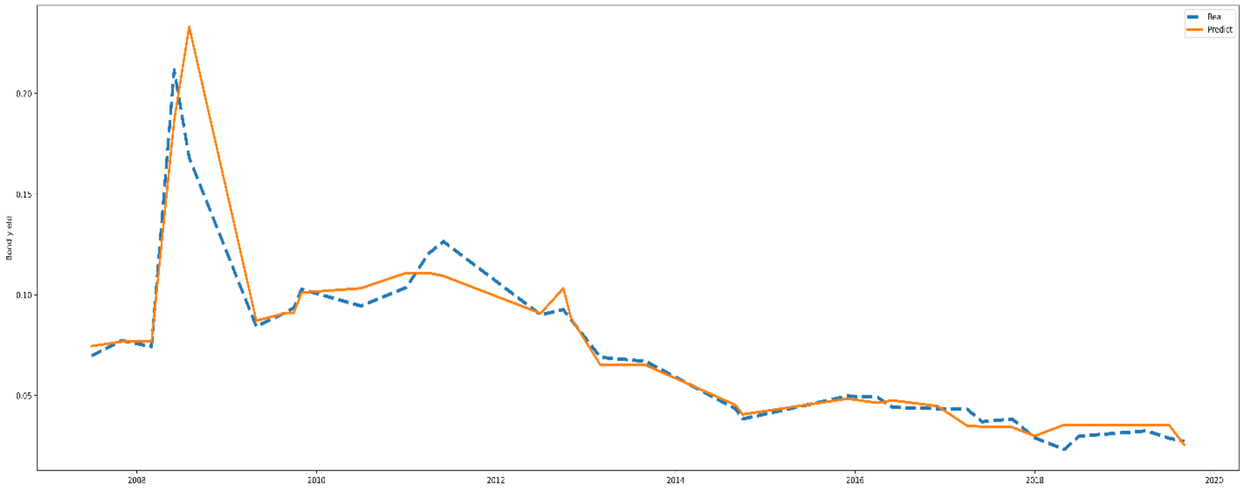

Figure 10. Actual and predicted 1-year government bond yields by Decision Tree

Source: Model results from Python

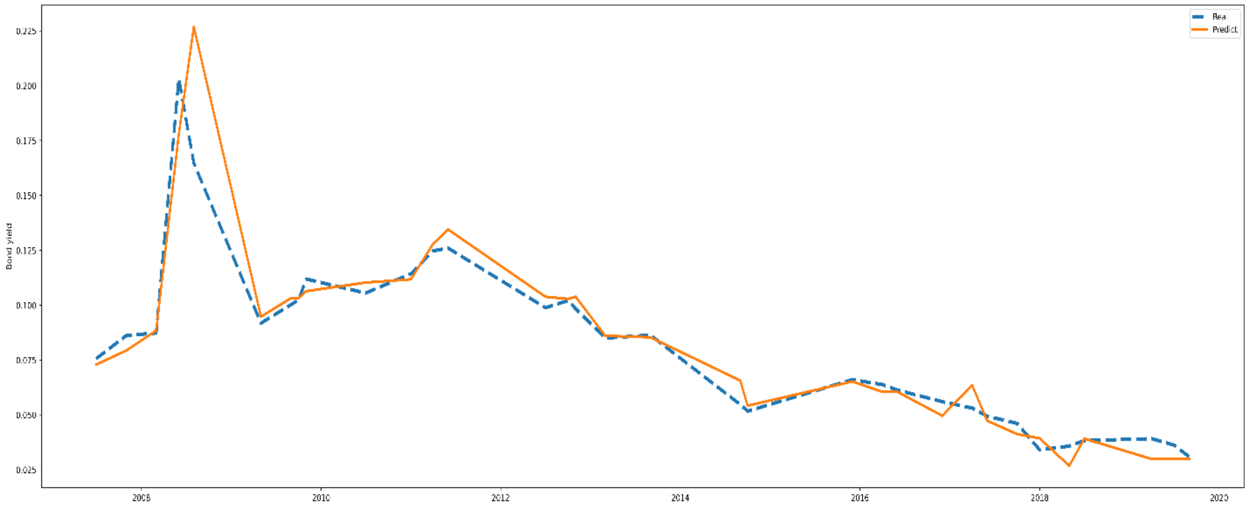

Figure 11. Actual and predicted 5-year government bond yields by Decision Tree

Source: Model results from Python

According to the forecast results of the Random Forest model in the short term and long term as shown in Figure 12 and Figure 13, we can see that the forecast line is closer to the actual line. More specifically, in Table 4, the 1-year forecast results have an error index of RMSE of 0.006858 and MAPE of 7.12%. As for the 5-year forecast results, the forecast results have RMSE and MAPE errors of 0.006185 and 4.34%, respectively. This shows that the accuracy of this model is higher than the Decision Tree model.

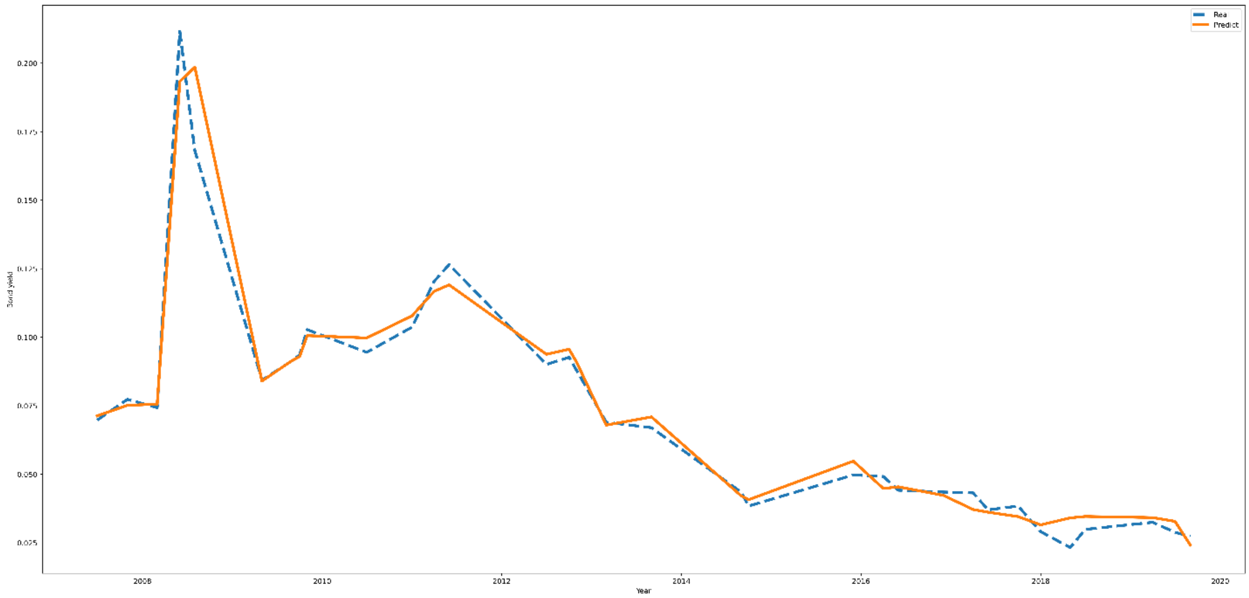

Figure 12. Actual and predicted 1-year government bond yields by Random Forest

Source: Model results from Python

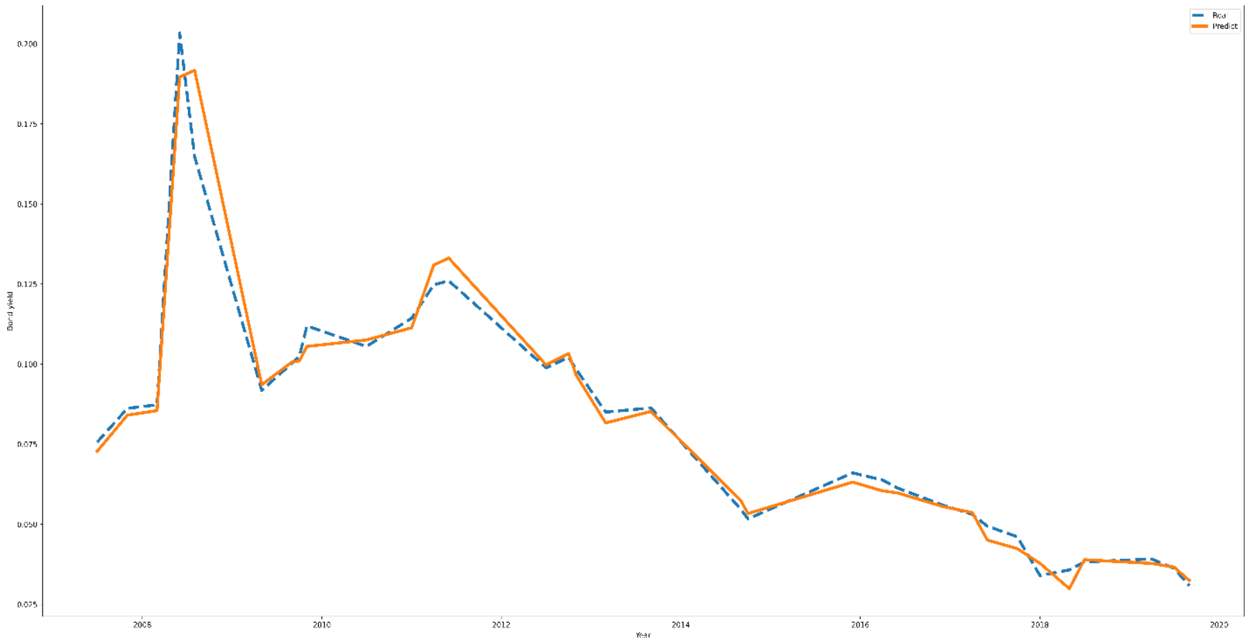

Figure 13. Actual and predicted 5-year government bond yields by Random Forest

Source: Model results from Python

In summary, the Random Forest model has solved the problem that the Decision Tree has not been optimal in minimizing the error in the residuals. The authors' results are also similar to previous studies on the topic of comparing predictive performance between these two models (Ganguli & Dunnmon, 2017). However, the two Machine Learning methods Decision Tree and Random Forest have many similarities as well as the operation relatively similar (Götze, Gürtler, & Witowski, 2020). The Random Forest model was developed based on the Decision Tree model with the aim of improving the predictive performance of the model itself (Götze et al., 2020). In addition, another Machine Learning model, SVR, has the outstanding advantages of excellent generalization ability, with high prediction accuracy. It is in contrast to traditional regression models that try to minimize the error of the coefficients of the equation in the model instead of trying to minimize the squared error in order to find a hyperplane that can fit the most data. Hence, the authors continue to make predictions using the SVR model.

Through Figure 14 and Figure 15, the shape of the kernel function directly affects the values obtained from the SVR model and gives different errors. The results of measuring the deviation obtained after applying the three kernel parameters of the SVR model are recorded in Table 4. These results are also measured by the RMSE, and MAPE indexes to evaluate the efficiency of the different kernels. When comparing the outputs for 1-year government bonds and 5-year government bonds, it can be seen that a decreasing value of deviation from the forecast is found in the use of linear and RBF kernels. Furthermore, the errors measured by MAPE were reduced when the linear kernel was used.

In addition, because of the local nature of the RBF kernel, it has superior learning ability compared to the rest of the kernels but lacks good model generalization, which is highly likely to lead to cases where overfitting or underfitting occurs in Machine Learning, reducing the accuracy of the predictive model. Therefore, the authors expect to find the optimal kernel parameter with high generalizability among the three used kernel parameters in particular and to find the optimal predictive model in general.

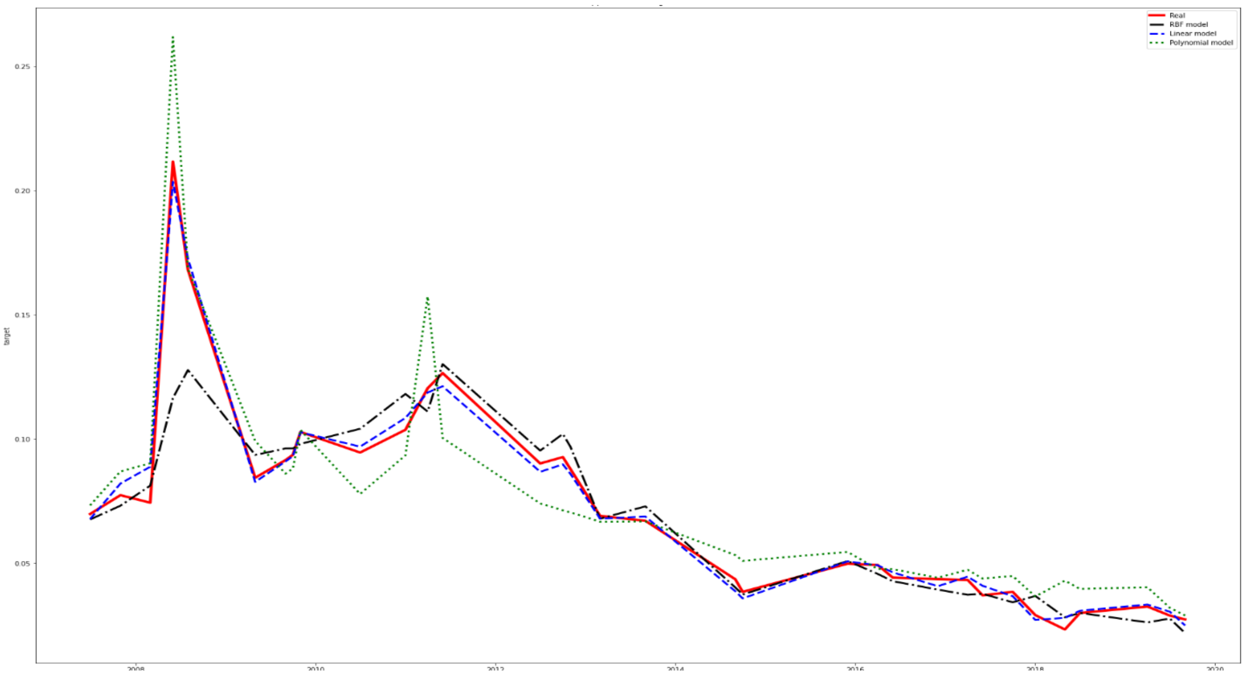

Figure 14. Actual and predicted 1-year government bond yields by SVR

Source: Model results from Python

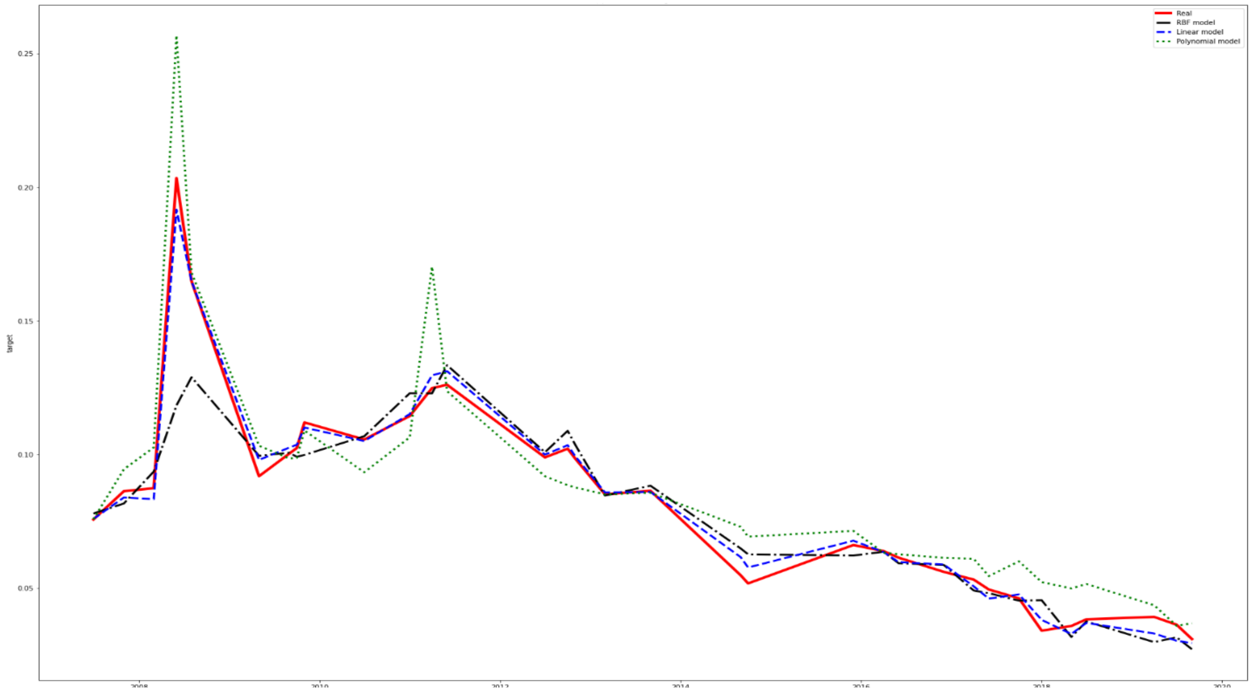

Figure 15. Actual and predicted 5-year government bond yields by SVR

Source: Model results from Python

Through the calculation results, the authors conclude that using SVR with the linear kernel has better predictive power than the remaining kernels, meeting the author's desire for a good kernel. In particular, it works more effectively for 5-year government bonds. To consider the significance of the results, we need to compare these results with the bias values generated by the traditional models and also the Machine Learning models implemented in this article. Specifically, based on Table 4, for 1-year or 5-year government bonds, the RMSE and MAPE indexes of the traditional models are higher than those of the SVR model in particular and other machine learning models tested in this article in general. This shows that there are still limitations in the reliability and accuracy of traditional models even if these traditional models give forecasts for 5-year government bonds that are different from the real ones have been minimized. Notably, although the Random Forest model has superior predictive power compared to the Decision Tree model, the Support Vector Regression model with the kernel parameters is passed as the basis RBF kernel and polynomial kernel. However, when compared with the Support Vector Regression with linear kernel, Random Forest still gives the results with somewhat weaker model accuracy even though its MAPE index is better than in the Government bond 5-year term. Specifically, based on Table 4, Random Forest's RMSE and MAPE are 0.006858 and 7.12%, respectively, for a 1-year term; and for a 5-year term, it is 0.006185 and 4.34%, respectively. These values are still high compared to Linear Kernel Support Vector Regression when its RMSE and MAPE are 0.003933 and 4.85% at 1-year term, respectively; 0.003818 and 4.52% at 5-year maturity.

In addition, it is noted that although the MAPE index given in 5-year government bonds of the Random Forest model (4.34%) is smaller than that of the Linear kernel Support Vector Regression (4.52%). However, this difference does not affect the selection of the more optimal model. Because in general, the Support Vector Regression model with linear kernels retains more stability and accuracy for both maturities of Government bonds than the Random Forest model.

V. CONCLUSION & RECOMMENDATIONS

Conclusion

Through the forecast results provided by the models, both 1-year- and 5-year government bond yields, the Support Vector Regression model shows that the difference to the reality is lowest and most stable. Besides, the research results also show that the 5-year government bond yield forecast always returns better results than the 1-year term. This shows that the long-term forecast for government bond yields is clearer, while the short-term forecast is more random.

Traditional models such as ARIMA, multivariable linear regression do well in forecasting the volatility trend of government bond yields. Besides, they also show the correlation relationship between past government bond yields and future yields through the ARIMA model or the correlations between macro variables and bond yields in a traditional linear regression model. However, this model has no advantage in giving high prediction accuracy, they still have relative deviation when compared with actual results. These models still have inherent limitations such as having to use many assumptions and not optimizing the residuals of the model. As for Machine Learning models, there will often be no assumptions, but mainly implementation of the "trial - error" method and will not care about the correlation relationship between the variables, which greatly improves the accuracy of the forecast.

Recommendations

Through the results obtained after implementing forecasting models, it can be seen that when combined with macro variables, the model is capable of giving good forecasting results of government bond yields. This is shown more clearly when implementing on a multivariable regression model, it will give more positive predictive results than when using a single time series variable with the ARIMA model. Besides, this study has contributed to reinforce the view that Machine Learning models will provide better predictive performance than traditional models. However, this method only brings high value to investors, and does not mean too much to policy makers, economic managers or State agencies. What policy makers are interested in is clarifying the effects of macro variables on government bond yields so that they can make reasonable policies and adjustments to achieve their expected economic goals. Therefore, depending on the purpose of use, analysts need to consider choosing an appropriate model for the decision-making process.

Limitations and recommendations for future research

Besides the results obtained from the above study by the authors, this study still has some limitations in the research process as follows:

- Firstly, the authors only use data in Vietnam for the period 2006 to 2019 to evaluate the forecast of government bond yields due to the limited of availability data resources. Therefore, in the next research, the authors will try to expand the data set with a longer time series, close to the current, or expand the scope of the study, such as studying government bond yields in countries around Southeast Asia, South Asia, etc. for giving the higher research value.
- Second, because of the limitation of knowledge and research period in the field of Machine Learning, the authors only use popular Machine Learning models such as Decision Tree, Random Forest, and Support Vector Regression at the simple application level. The models used by the authors have not yet had steps to calibrate the parameters inside the model's algorithm. This study will be the foundation for the authors to develop predictive research from a more advanced level.
- Third, the authors have not completely found the most optimal macro variables for forecasting short-term government bond yields. Besides, this study does not has practical factors and market sensibility such as daily market news, transaction costs, politics, etc. having a more objective view in assessing the importance of government bond yields as well as supporting forecast.

REFERENCES

[1] Agnihotri, A. (2015). Study of various factors affecting return in bond market-A case of India. *AIMA Journal of Management and Research, 9*(2).

[2] Akram, T., & Das, A. (2019). The long-run determinants of Indian government bond yields. *Asian Development Review, 36*(1), 168-205.

[3] Ang, A., & Piazzesi, M. (2003). A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. *Journal of Monetary economics, 50*(4), 745-787.

[4] Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics, 11*, 685-725.
[5] Attigeri, G. V., MM, M. P., Pai, R. M., & Nayak, A. (2015). *Stock market prediction: A big data approach.* Paper presented at the TENCON 2015-2015 IEEE Region 10 Conference.
[6] Bansal, R., & Shaliastovich, I. (2013). A long-run risks explanation of predictability puzzles in bond and currency markets. *The Review of Financial Studies, 26*(1), 1-33.
[7] Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies, 34*(2), 1046-1089.
[8] Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.
[9] Campbell, J. Y., & Shiller, R. J. (1991). Yield spreads and interest rate movements: A bird's eye view. *The Review of Economic Studies, 58*(3), 495-514.
[10] Chen, L., Pelger, M., & Zhu, J. (2020). Deep learning in asset pricing. *Available at SSRN 3350138*.
[11] Cherif, P. M., & Kamoun, S. (2007). TERM STRUCTURE DYNAMIC AND MACROECONOMIC FUNDAMENTALS IN THE EURO AREA. *International Journal of Finance, 19*(3).
[12] Cieslak, A., & Povala, P. (2015). Expected returns in Treasury bonds. *The Review of Financial Studies, 28*(10), 2859-2901.
[13] Cooper, I., & Priestley, R. (2009). Time-varying risk premiums and the output gap. *The Review of Financial Studies, 22*(7), 2801-2833.
[14] Farrell, M. H., Liang, T., & Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica, 89*(1), 181-213.
[15] Feng, G., He, J., & Polson, N. G. (2018). Deep learning for predicting asset returns. *arXiv preprint arXiv:1804.09314*.
[16] Feng, G., Polson, N. G., & Xu, J. (2018). Deep Learning in Characteristics-Sorted Factor Models. *arXiv preprint arXiv:1805.01104*.
[17] Gan, L., Wang, H., & Yang, Z. (2020). Machine learning solutions to challenges in finance: An application to the pricing of financial products. *Technological Forecasting and Social Change, 153*, 119928.
[18] Ganguli, S., & Dunnmon, J. (2017). Machine learning for better models for predicting bond prices. *arXiv preprint arXiv:1705.01142*.
[19] Ghysels, E., Horan, C., & Moench, E. (2018). Forecasting through the rearview mirror: Data revisions and bond return predictability. *The Review of Financial Studies, 31*(2), 678-714.
[20] Gilles, C., & LeRoy, S. F. (1991). On the arbitrage pricing theory. *Economic theory, 1*(3), 213-229.
[21] Giordano, L., Linciano, N., & Soccorso, P. (2012). The determinants of government yield spreads in the euro area.
[22] Götze, T., Gürtler, M., & Witowski, E. (2020). Improving CAT bond pricing models via machine learning. *Journal of Asset Management, 21*(5), 428-446.
[23] Greenwood, R., & Vayanos, D. (2014). Bond supply and excess bond returns. *The Review of Financial Studies, 27*(3), 663-713.
[24] Gu, S., Kelly, B., & Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics, 222*(1), 429-450.
[25] Han, S.-H., Kang, K., & Shin, Y. S. (2016). Bond ratings, corporate governance, and cost of debt: The case of Korea. *The Journal of Asian Finance, Economics, and Business, 3*(3), 5-15.
[26] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Second Edition ed.). Springer Science & Business Media.
[27] Heaton, J., Polson, N. G., & Witte, J. (2016). Deep portfolio theory. *arXiv preprint arXiv:1605.07230*.
[28] Hoogteijling, T. (2020). Forecasting Bond Risk Premia with Machine Learning.
[29] Huang, J.-z., & Shi, Z. (2011). *Determinants of bond risk premia.* Paper presented at the AFA Denver Meetings Paper.
[30] Inoguchi, M. (2007). Influence of ADB bond issues and US bonds on Asian Government bonds. *Asian Economic Journal, 21*(4), 387-404.
[31] Kameda, K. (2014). Budget deficits, government debt, and long-term interest rates in Japan. *Journal of the Japanese and International Economies, 32*, 105-124.

[32] Kim, S., & Kang, M. (2019). Financial series prediction using Attention LSTM. *arXiv preprint arXiv:1902.10877*.
[33] Kluza, S., & Sławiński, A. (2002). *FACTORS AFFECTING T-BOND PRICES: FROM INVESTORS PERSPECTIVE.* Paper presented at the 22nd NBP Conference "Monetary Policy in the environment of structural changes.
[34] Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research, 259*(2), 689-702.
[35] Kumar, M. M. S., & Baldacci, M. E. (2010). *Fiscal deficits, public debt, and sovereign bond yields*: International Monetary Fund.
[36] Ludvigson, S. C., & Ng, S. (2009). Macro factors in bond risk premia. *The Review of Financial Studies, 22*(12), 5027-5067.
[37] Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives, 31*(2), 87-106.
[38] Nguyễn Anh Phong, N. T. C., Phan Huy Tâm, Ngô Phú Thanh. (2020). *Sách tham khảo "Ứng dụng Python trong tài chính"*: NXB Đại học Quốc Gia Tp.HCM.
[39] NGUYEN, H. H. (2019). The role of state budget expenditure on economic growth: empirical study in Vietnam. *The Journal of Asian Finance, Economics and Business, 6*(3), 81-89.
[40] Olson, D., & Mossman, C. (2003). Neural network forecasts of Canadian stock returns using accounting ratios. *International Journal of Forecasting, 19*(3), 453-465.
[41] Peters, J., De Baets, B., Verhoest, N. E., Samson, R., Degroeve, S., De Becker, P., & Huybrechts, W. (2007). Random forests as a tool for ecohydrological distribution modelling. *ecological modelling, 207*(2-4), 304-318.
[42] Reschreiter, A. (2003). *Risk factors of inflation-indexed and conventional Government bonds and the APT.* Paper presented at the Money, Macro and Finance (MMF) Conference, University of Cambridge, UK.
[43] Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews, 71*, 804-818.
[44] Santosa, P. W., & Sihombing, P. (2015). Factors affecting the yeld curve fluctuation of Indonesia government bond. *International Journal of Applied Business and Economic Research, 13*(6), 4073-4087.
[45] Soni, S. (2011). Applications of ANNs in stock market prediction: a survey. *International Journal of Computer Science & Engineering Technology, 2*(3), 71-83.
[46] Suimon, Y. (2018). Fluctuation model of JGB yield curve using machine learning and the interest rate prediction. *The Japanese Society for Artificial Intelligence, SIG-FIN*(21), 46-49.
[47] Suimon, Y., Sakaji, H., Shimada, T., Izumi, K., & Matsushima, H. (2019). *Japanese long-term interest rate forecast considering the connection between the Japanese and US yield curve.* Paper presented at the 2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr).
[48] TRINH, Q. T., NGUYEN, A. P., NGUYEN, H. A., & NGO, P. T. (2020). Determinants of Vietnam government bond yield volatility: A GARCH approach. *The Journal of Asian Finance, Economics, and Business, 7*(7), 15-25.
[49] Ul Haque, N., Mark, N. C., & Mathieson, D. J. (1998). The relative importance of political and economic variables in creditworthiness ratings.
[50] Valko, M., Marques, N. C., & Castellani, M. (2005). *Evolutionary feature selection for spiking neural network pattern classifiers.* Paper presented at the 2005 portuguese conference on artificial intelligence.
[51] Wang, H., & Xu, D. (2017). Parameter selection method for support vector regression based on adaptive fusion of the mixed kernel function. *Journal of Control Science and Engineering, 2017.*
[52] Yieand, C., & Chen, N. (2019). Determinants of bond yield. *International Journal of Recent Technology and Engineering (IJRTE), 7*, 238-244.
[53] Zaja, M. M., Jakovcevic, D., & Visic, L. (2018). Determinants of the government bond yield: Evidence from a highly euroised small open economy. *International Journal of Economic Sciences, 7*(2), 87-106.

APPENDIX

Appendix Table 1: Forecast results by ARIMA

Date	Government bond yield 1 year		Government bond yield 5 year	
	Real	**Forcast**	**Real**	**Forcast**
2017-05	0.038	0.043887759	0.05078	0.053156275
2017-06	0.037	0.047592825	0.0494	0.056894996
2017-07	0.0336	0.045908238	0.04712	0.056909442
2017-08	0.0411	0.044108181	0.04986	0.054934579
2017-09	0.04	0.044026615	0.04708	0.053043346
2017-10	0.0384	0.043156971	0.0461	0.052242871
2017-11	0.04054	0.042832672	0.04636	0.052174557
2017-12	0.0364	0.042776115	0.0436	0.051891294
2018-01	0.02902	0.040474959	0.034	0.049154181
2018-02	0.02575	0.039289503	0.03363	0.047845274
2018-03	0.0236	0.037549293	0.03143	0.046112685
2018-04	0.025	0.039088072	0.03613	0.047387081
2018-05	0.02325	0.041134404	0.03575	0.048431522
2018-06	0.02338	0.044839469	0.0365	0.051812111
2018-07	0.02988	0.043154883	0.03825	0.05190076
2018-08	0.0422	0.041354826	0.04795	0.049821577
2018-09	0.04088	0.04127326	0.04458	0.047403366
2018-10	0.04275	0.040403616	0.04745	0.04664054
2018-11	0.0415	0.040079316	0.04625	0.046948899
2018-12	0.041	0.04002276	0.04555	0.046928497
2019-01	0.03362	0.037721604	0.03868	0.044185699
2019-02	0.031	0.036536148	0.0378	0.042726733
2019-03	0.03257	0.034795937	0.03878	0.04143821
2019-04	0.03247	0.036334716	0.03915	0.042655296
2019-05	0.03255	0.038381049	0.03845	0.04353421
2019-06	0.03273	0.042086114	0.0391	0.046885997
2019-07	0.02883	0.040401527	0.03622	0.046969635
2019-08	0.028	0.038601471	0.03318	0.04488958
2019-09	0.02729	0.038519904	0.03086	0.042471217
2019-10	0.02343	0.03765026	0.02672	0.041708365
2019-11	0.0215	0.037325961	0.02405	0.042016719
2019-12	0.0155	0.037269404	0.02	0.041996317
Model results from Python				

Appendix Table 2: Forecast results by multi-variable regression

Date	Government bond yield 1 year		Government bond yield 5 year	
	Real	Forcast	Real	Forcast
2017-05	0.038	0.03856	0.05078	0.04806
2017-06	0.037	0.03667	0.0494	0.0478
2017-07	0.0336	0.03345	0.04712	0.04557
2017-08	0.0411	0.03556	0.04986	0.05256
2017-09	0.04	0.03291	0.04708	0.05121
2017-10	0.0384	0.03207	0.0461	0.0497
2017-11	0.04054	0.03306	0.04636	0.05085
2017-12	0.0364	0.02946	0.0436	0.04786
2018-01	0.02902	0.02027	0.034	0.04049
2018-02	0.02575	0.01912	0.03363	0.03841
2018-03	0.0236	0.01797	0.03143	0.03524
2018-04	0.025	0.02213	0.03613	0.03693
2018-05	0.02325	0.02127	0.03575	0.03622
2018-06	0.02338	0.02229	0.0365	0.03637
2018-07	0.02988	0.02499	0.03825	0.0403
2018-08	0.0422	0.03546	0.04795	0.0506
2018-09	0.04088	0.03146	0.04458	0.04978
2018-10	0.04275	0.03437	0.04745	0.05091
2018-11	0.0415	0.03419	0.04625	0.04893
2018-12	0.041	0.03317	0.04555	0.04864
2019-01	0.03362	0.02633	0.03868	0.03941
2019-02	0.031	0.02466	0.0378	0.03838
2019-03	0.03257	0.0268	0.03878	0.03813
2019-04	0.03247	0.02653	0.03915	0.03885
2019-05	0.03255	0.02545	0.03845	0.03919
2019-06	0.03273	0.02691	0.0391	0.03847
2019-07	0.02883	0.02355	0.03622	0.03579
2019-08	0.028	0.01994	0.03318	0.0354
2019-09	0.02729	0.01734	0.03086	0.03489
2019-10	0.02343	0.01245	0.02672	0.03236
2019-11	0.0215	0.009	0.02405	0.03151
2019-12	0.0155	0.00397	0.02	0.0275
				Model results from Python

Appendix Table 3: Forecast results by Decision Tree

Date	Government bond yield 1 year		Government bond yield 5 year	
	Real	Forcast	Real	Forcast
2007-07	0.06975	0.074544443	0.07559	0.07292548
2007-11	0.07733	0.076783919	0.08617	0.07934872
2008-03	0.0743	0.076820235	0.0873	0.088419708

2008-06	0.21167	0.186578767	0.20333	0.177977324
2008-08	0.16814	0.233123551	0.16475	0.226887109
2009-05	0.08435	0.08715814	0.09178	0.094757141
2009-09	0.09145	0.091043934	0.10028	0.103251998
2009-10	0.09358	0.091043934	0.10233	0.103251998
2009-11	0.10278	0.101115523	0.11191	0.106365554
2010-07	0.0945	0.103294472	0.10551	0.110300401
2011-01	0.10367	0.110908691	0.11431	0.111771372
2011-04	0.12038	0.110908691	0.12473	0.127927536
2011-06	0.1265	0.109407636	0.126	0.134510132
2012-07	0.0901	0.090813933	0.09883	0.103815871
2012-10	0.09267	0.103294472	0.10208	0.102847481
2012-11	0.08783	0.088634984	0.09833	0.103815871
2013-03	0.069	0.065162855	0.085	0.086041638
2013-09	0.067	0.065162855	0.08633	0.085342927
2014-09	0.04346	0.045334414	0.05472	0.065631916
2014-10	0.0384	0.040649672	0.05168	0.0541706
2015-12	0.04974	0.048457575	0.06604	0.065202883
2016-04	0.0492	0.046375467	0.06378	0.060483517
2016-06	0.04416	0.047610205	0.0613	0.060483517
2016-12	0.0435	0.044789676	0.056	0.049573816
2017-04	0.0432	0.034996508	0.0531	0.063584814
2017-06	0.037	0.034597034	0.0494	0.047306069
2017-10	0.0384	0.034294402	0.0461	0.041238314
2018-01	0.02902	0.029936503	0.034	0.039289277
2018-05	0.02325	0.03550493	0.03575	0.02674925
2018-07	0.02988	0.03550493	0.03825	0.039154438
2019-04	0.03247	0.035359667	0.03915	0.029875063
2019-07	0.02883	0.035359667	0.03622	0.029875063
2019-09	0.02729	0.025493867	0.03086	0.029875063
				Model results from Python

Appendix Table 4: Forecast results by Random Forest

Date	Government bond yield 1 year		Government bond yield 5 year	
	Real	**Forcast**	**Real**	**Forcast**
2007-07	0.06975	0.071314514	0.07559	0.07274737
2007-11	0.07733	0.075161086	0.08617	0.084057788
2008-03	0.0743	0.075599539	0.0873	0.085474579
2008-06	0.21167	0.193344284	0.20333	0.189631949
2008-08	0.16814	0.198551369	0.16475	0.191697683
2009-05	0.08435	0.084011979	0.09178	0.093553642
2009-09	0.09145	0.09151265	0.10028	0.100637715
2009-10	0.09358	0.092920372	0.10233	0.101003987
2009-11	0.10278	0.100604317	0.11191	0.105508836

2010-07	0.0945	0.099750774	0.10551	0.10754184
2011-01	0.10367	0.107796303	0.11431	0.111325423
2011-04	0.12038	0.11673956	0.12473	0.13091373
2011-06	0.1265	0.119121878	0.126	0.133102167
2012-07	0.0901	0.093779605	0.09883	0.0997593
2012-10	0.09267	0.095634617	0.10208	0.103264869
2012-11	0.08783	0.090833907	0.09833	0.096480996
2013-03	0.069	0.067952637	0.085	0.081618796
2013-09	0.067	0.070924361	0.08633	0.085234075
2014-09	0.04346	0.042109205	0.05472	0.057494504
2014-10	0.0384	0.040713104	0.05168	0.053355682
2015-12	0.04974	0.05479856	0.06604	0.063159949
2016-04	0.0492	0.044804929	0.06378	0.060415485
2016-06	0.04416	0.045413461	0.0613	0.059770342
2016-12	0.0435	0.042303253	0.056	0.055553803
2017-04	0.0432	0.037137811	0.0531	0.053631244
2017-06	0.037	0.03614772	0.0494	0.045057812
2017-10	0.0384	0.034547887	0.0461	0.042394374
2018-01	0.02902	0.031537426	0.034	0.037818674
2018-05	0.02325	0.03408026	0.03575	0.029907547
2018-07	0.02988	0.034603934	0.03825	0.038930851
2019-04	0.03247	0.034185092	0.03915	0.037766209
2019-07	0.02883	0.032816348	0.03622	0.036630375
2019-09	0.02729	0.024186376	0.03086	0.032439333
				Model results from Python

Appendix Table 5: Forecast results by SVR

Government bond yield 1 year				
Date	**Real**	**RBF**	**Linear**	**Polynomial**
2007-07	0.06975	0.067546791	0.067725197	0.073243069
2007-11	0.07733	0.073086109	0.082010615	0.08689592
2008-03	0.0743	0.081174462	0.088768768	0.090195361
2008-06	0.21167	0.116417177	0.203551541	0.261829382
2008-08	0.16814	0.127730556	0.17262905	0.168581713
2009-05	0.08435	0.093523938	0.082659023	0.0991682
2009-09	0.09145	0.096134574	0.090905907	0.086007949
2009-10	0.09358	0.096164624	0.093048956	0.08803777
2009-11	0.10278	0.09797871	0.102664244	0.103575093
2010-07	0.0945	0.104071805	0.096866041	0.077809505
2011-01	0.10367	0.118100348	0.108437475	0.093376658
2011-04	0.12038	0.111020804	0.118670691	0.157344363
2011-06	0.1265	0.130116986	0.12119283	0.100382147
2012-07	0.0901	0.095286439	0.086746147	0.07398246

2012-10	0.09267	0.101989319	0.089733353	0.071219275
2012-11	0.08783	0.096783137	0.085889452	0.070391568
2013-03	0.069	0.067853912	0.067898462	0.066651895
2013-09	0.067	0.072822267	0.068700486	0.066781493
2014-09	0.04346	0.040170967	0.038638244	0.053213214
2014-10	0.0384	0.037298205	0.035748925	0.050887674
2015-12	0.04974	0.050818655	0.050661252	0.054470804
2016-04	0.0492	0.045618984	0.049090873	0.04757547
2016-06	0.04416	0.04268589	0.046253267	0.047542764
2016-12	0.0435	0.039297215	0.040666025	0.044070344
2017-04	0.0432	0.037207415	0.044488344	0.047297073
2017-06	0.037	0.03760305	0.040827142	0.043752421
2017-10	0.0384	0.034194209	0.03664173	0.044777722
2018-01	0.02902	0.036726903	0.027073171	0.036525369
2018-05	0.02325	0.02815737	0.027926918	0.042921552
2018-07	0.02988	0.029907986	0.03078229	0.039570154
2019-04	0.03247	0.026090864	0.033290444	0.040268973
2019-07	0.02883	0.027635584	0.03038692	0.031958836
2019-09	0.02729	0.021780927	0.02486533	0.028898418

Government bond yield 5 year

Date	Real	RBF	Linear	Polynomial
2007-07	0.07559	0.077848899	0.075814819	0.075695254
2007-11	0.08617	0.081567548	0.083812231	0.094368566
2008-03	0.0873	0.093514039	0.083221333	0.102540735
2008-06	0.20333	0.118271523	0.191542363	0.25667375
2008-08	0.16475	0.128824199	0.164747386	0.168005626
2009-05	0.09178	0.099438003	0.097880807	0.103192583
2009-09	0.10028	0.100571446	0.102509853	0.098402512
2009-10	0.10233	0.099035273	0.103547798	0.09864959
2009-11	0.11191	0.099767507	0.109992232	0.10886412
2010-07	0.10551	0.106755128	0.104949406	0.093252169
2011-01	0.11431	0.122839365	0.11502652	0.106329864
2011-04	0.12473	0.122789415	0.129503121	0.170130058
2011-06	0.126	0.133325353	0.131079731	0.123585118
2012-07	0.09883	0.100754545	0.099873058	0.091781684
2012-10	0.10208	0.108780133	0.103385821	0.088391845
2012-11	0.09833	0.103487475	0.099208211	0.087582086
2013-03	0.085	0.084534658	0.085705906	0.084980375
2013-09	0.08633	0.088237307	0.085913226	0.085602722
2014-09	0.05472	0.064756369	0.061513969	0.072862274
2014-10	0.05168	0.062570025	0.057634251	0.069165884
2015-12	0.06604	0.062102732	0.067668111	0.071342287
2016-04	0.06378	0.063511623	0.06341037	0.063179517
2016-06	0.0613	0.059225799	0.059648358	0.062586537

2016-12	0.056	0.058578543	0.0587559	0.061241105
2017-04	0.0531	0.049012994	0.050578128	0.060898061
2017-06	0.0494	0.04802658	0.045983938	0.054424153
2017-10	0.0461	0.045224061	0.047546992	0.059919177
2018-01	0.034	0.045380812	0.038028454	0.052166504
2018-05	0.03575	0.031651874	0.032747299	0.04980257
2018-07	0.03825	0.037449622	0.036922337	0.051456774
2019-04	0.03915	0.02970815	0.032926683	0.043499227
2019-07	0.03622	0.031511654	0.030225839	0.035827794
2019-09	0.03086	0.027036894	0.029344163	0.036759372
				Model results from Python