Norwegian School of Economics
Bergen, Fall 2019

Forecasting GDP Growth

A Comprehensive Comparison of Employing Machine Learning Algorithms and Time Series Regression Models

Pirasant Premraj

Supervisor: Jonas Andersson

Master thesis, Economics and Business Administration

Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

This page is intentionally left blank.

Acknowledgements

This thesis is the second last chapter of my Master of Science (M.Sc) degree at the Norwegian School of Economics (NHH). It was written during the autumn of 2019, in accordance with my specialisation in Business Analytics (BAN).

I would first thank my supervisor Jonas Andersson for his guidance and excellent input. I would also thank Thomas Hansen and Nora Hansen from PwC for their advice, contribution and pointers on the methodology and the analysis which has lifted the quality of the thesis.

Last but not least, I would also thank my good friend Johannes Tyrihjell, for valuable discussions during my work with this thesis.

Pirasant Premraj,

Oslo, December 19, 2019

Abstract

In this paper, we do a comprehensive comparison of forecasting Gross Domestic Product (GDP) growth using Machine Learning algorithms and traditional time series regression models on the following economies: *Australia, Canada, Euro Area, Germany, Spain, France, Japan, Sweden, Great Britain* and *USA*. The ML algorithms we employ are Bayesian Additive Trees Regression Trees (BART), Elastic-Net Regularized Generalized Linear Models (GLMNET), Stochastic Gradient Boosting (GBM) and eXtreme Gradient Boosting (XGBoost), while Autoregressive (AR) models, Autoregressive Integrated Moving Average (ARIMA) models and Vector Autoregressive (VAR) models represents the traditional time series regression methods. The results assert that the multivariate VAR models are superior, indicating the chosen variables' and the models' suitability of forecasting GDP growth. Furthermore, we also do an assessment of the top three variables that drives the best performing Machine Learning algorithm of XGBoost to investigate whether it suggests the same variables in forecasting GDP growth as macroeconomic theory. In general we do see some evidence, but in many cases the algorithm emphasizes other variables than what macroeconomic theory suggests.

Keywords – Time Series, Machine Learning, Econometric, GDP, Forecast

Abbreviations

AR Autoregressive Model

ARIMA Autoregressive Integrated Moving Average Model

BART Bayesian Additive Regression Trees

GBM Stochastic Gradient Boosting Machine

GDP Gross Domestic Product

GLMNET Elastic-Net Regularized Generalized Linear Model

ML Machine Learning

TS Time Series

VAR Vector Autoregressive Model

XGBoost eXtreme Gradient Boosting

Contents

1 Introduction

Policy makers need to understand the state of the economy in order to make the best possible policy decisions. These decisions are often made under uncertainty not only regarding future economic conditions, but also regarding the current economic situation. Key macroeconomic statistics are often released with lags and are subject to frequent revisions. This has led to many central banks and institutions building forecasting models to mitigate these uncertainties and obtain a timelier, yet accurate indication of the state of the economy.

Generally, the forecasting models have been built on the basis of time series regression, while there has been little progress in employing ML algorithms to forecast macroeconomic variables. ML algorithms are able to handle large datasets and to detect underlying, complex relationships between variables. This ability has made ML algorithms widely used in different fields. In the health care industry ML is used to analyze patient health data and to flag anomalies and to detect warning signs which would not normally be detected, while in the financial industry ML is used to identify important insight in data, automate stock trade and prevent fraud. In addition, in the retail industry ML is used to give personalized product recommendations while adjusting price to match real time changes in demand. Despite ML algorithms not being conceptually different from other statistical models in terms of modelling an outcome y from a function f, little progress has been made in understanding the properties of ML and its application in forecasting macroeconomic outcomes. There are several possible explanations to this fact. Firstly, forecasting macroeconomic outcomes has typically been done in the field of econometrics, and thus it is easier to expand an existing framework rather than move to the new field of ML. Secondly, many economic applications revolve around parameter estimation and causal inference, while ML algorithms are not built for this purpose (Mullainathan and Spiess, 2017).

In current literature on the topic there is limited research on the use of ML algorithms to forecast the state of the economy and comparing the results to traditional time series regression techniques. This thesis aims to contribute to this field by giving insights on how ML algorithms perform on forecasting GDP growth compared to traditional time series regression models

1.1 Problem formulation

This thesis aims to forecast GDP growth of ten economies using ML algorithms, and compare these forecasts to those made by traditional forecasting models. The economies included for this research are: *Australia, Canada, Euro Area, Germany, Spain, France, Japan, Sweden, Great Britain* and *USA*. The forecast performance of the ML algorithms will not only be compared with each other, but also with traditional univariate and multivariate forecasting methods used by central banks and other institutions. By putting an emphasis on the comparison between ML algorithms and traditional methods, we wish to make this paper appealing and relevant to both academics and decision-makers in central banks and other relevant institutions. Based on our motivation and scope, the following research questions will be investigated:

1. How well do state-of-the-art ML algorithms perform on forecasting out-of-sample GDP growth and do these algorithms have the ability to outperform traditional univariate and multivariate forecasting models?
2. Do these ML algorithms suggest other predictors for forecasting GDP growth than what economic theory suggests?

By exploring these topics, we wish to investigate whether employing this technology delivers better forecast performance than traditional time series techniques, and its potential to be used by decision-makers in central banks and other relevant institutions. Furthermore, we also want to contrast ML with economic theory to see whether the ML methods suggest other predictors for GDP forecasting. How our models would be implemented for real time GDP growth forecast, is beyond the scope of this thesis, as the implementation would have been too extensive. Moreover, we have not assessed the use of lagged variables in terms of shifting observations back in time, as we want to utilize the most recent observations in the forecasting. This thesis focuses on both the technical and the practical aspects of GDP growth forecasting, meaning that the technical foundation of the models are elaborated, while there is substantial emphasis on the achieved results.

1.2 Literature review

Employing ML algorithms in the forecasting of macroeconomic variables is a relatively new and growing field. However, most of the contributed academic research share many shortcomings among them: some focus on one particular ML algorithm on a limited time span, others only incorporate a few variables in developing the ML models and others only compare the results with very basic forecasting models. In this section we will summarize the main findings in the field, and elaborate on how our research differ from previous research on the topic.

Saman (2011) modelled various scenarios of the GDP development of Romania using neural network ML algorithms. The focus of this paper was to forecast GDP development by studying the non-linear relationship between GDP and investments, including domestic and direct investments. The research proposed two models, and showed that the models at least had the ability to forecast structural breaks in GDP.

Similarly, Tkacz (2001) applied neural network algorithms to forecast Canadian GDP growth. Although the research indicated that the neural network models yielded lower forecast errors on forecasting year-on-year GDP growth compared to traditional methods such as linear and univariate forecasting methods, the results indicated that the forecast improvements were less notable when forecasting quarterly GDP growth. Interestingly, the neural network models were also not able to outperform a naïve no change-model, which is a model anticipating that the level of the variable in the current period is the same as previous period. The conclusion of the research was that neural network models were probably more suitable in forecasting long-term GDP growth rather than short-term GDP growth.

The most related research on forecasting GDP growth using ML algorithms are done by Richardson et al. (2018) and Jung et al. (2018). Richardson et al. (2018) examined whether ML algorithms could improve forecasts of real GDP growth in New Zealand. They found that ML algorithms outperformed classic statistical methods, indicating the suitability of Support-Vector Machine (SVM), Neural Network (NN), Lasso, Boosted Tree (BT), Regularized Generalized Linear Model (GLMNET) and Ridge to forecast GDP growth. The International Monetary Fund (IMF) by Jung et al. (2018) published a working paper where they researched the application of ML algorithms to forecast the GDP growth of the advanced G7 economies and some emerging economies. While they employed the ML algorithms of GLMNET, NN, Super Learner, they also reached the conclusion that these algorithms outperformed standard classic statistical methods. Despite their findings, Richardson et al. (2018) and Jung et al. (2018) share two limitations. Firstly, both of these studies only use standard tuning parameters in order to fit the ML models. As the tuning parameters effect the learning ability of the ML models, these parameters should be carefully chosen. Secondly, the evaluation of forecast performance was only drawn to the accuracy measure of RMSE. One drawback of only using this accuracy measure, is that one single bad forecast point will skew the metric towards underestimating any model's suitability. In addition, Jung et al. (2018) do not include the plots of the forecasted values versus the real values of all the economies studied or a specified list of which variables that are used in building the ML algorithms. Ideally, for increased validity and transparency any reader should have access to a list of the all of the variables included and the plots in order to have the opportunity to assess the generalization of employing ML algorithms on GDP forecasting.

Based on the previously outlined motivation and the research done within the field, our contribution to the literature will be two-folded:

1. Firstly, we will contribute to existing research on using ML algorithms to forecast GDP growth using other ML algorithms than previous research. While there have been done research on some of the our chosen economies, it is therefore interesting to see if there are differences in forecast performance. We will go further than previous research by comparing the results of the ML algorithms to TS regression models, and also evaluate model's ability to factor in short-term fluctuations using both forecast accuracy measures and plots.

2. Secondly, we will compare the variables that drives the best performing ML algorithm with economic theory to see whether the ML algorithms suggest any interesting uncommon predictors.

1.3 Thesis structure

In chapter 2 we will start by briefly elaborating on the relevant macroeconomic theory and technical theory. Then, chapter 3 will explain the choice of data, whereas the chapter 4 outlines the theoretical methodology underlying the forecast models, how we will assess the forecast performance and the methodology behind extracting the most important variables from any of the chosen ML algorithm. In chapter 5 we will do an assessment of the results from the forecast models and examine the suggestions from the best performing ML algorithm on the most important variables that drives the forecasting model. Furthermore, in chapter 6 we will discuss the applicability of employing ML algorithms on GDP forecasting and share some thoughts on GDP forecasting in general. Finally, in chapter 7 we will make a conclusion.

2 Theory

In this chapter we will start by elaborating on GDP and its workings, before briefly looking into the workings of the chosen ML algorithms and the TS regression models.

2.1 Gross Domestic Product - GDP

One of the key measures of the state of the economy is Gross Domestic Product (GDP). GDP is defined as the market value of the final goods and services provided in an economy over a certain period, often annually or quarterly (Jones, 2014). There are two ways of noting the GDP, nominal and real GDP. GDP is normally collected at nominal level, and to compare the GDP across periods the nominal GDP needs to be adjusted for inflation. In that way we can determine whether an increase in GDP is due to increased production or just increased prices.

Theoretically, GDP can be calculated in three different ways where the production, income and expenditure approaches all amounts to the same value:

1. **Production approach**: adds the gross output in different industries and subtracts intermediate outputs. This difference represents the value added and prevents double counting.

2. **Income approach**: measures the income earned by different factors of production, by adding up all the income earned in the economy.

3. **Expenditure approach**: divides the goods and services that are purchased into several categories, a breakdown resulting in an equation called national income identity. The equation states that goods and services can be consumed, invested by the private sector, bought by the government or shipped abroad for foreigners to use. The equation is given by:

$$Y = C + I + G + NX \tag{2.1}$$

where $Y = GDP$, $C = Consumption$, $I = Investment$, $G = Government\ Purchases$ and $NX = Net\ Export = Export - Import$

GDP within countries are usually calculated by national statistics agencies, by gathering information from a wide range of sources. The calculations of GDP often follow international established standards contained in the System of National Account (Nations, 2010) developed by International Monetary Fund (IMF), the European Commission (EC), the Organization for Economic Cooperation and Development (OECD), the United Nations (UN) and the World Bank.

There are several factors that are not accounted for in calculating GDP. Firstly, the output produced by a country's citizens abroad is not taken into consideration, alongside with profits earned by companies outside their home country. This is however calculated in Gross National Product (GNP) as the market value of the final goods and services provides in an economy over a certain period of a country‘s citizens. Secondly, GDP only includes goods and services that are transacted in the market. The implication is that an amount spent at a restaurant will count positively towards GDP, while if one uses the same ingredients to a home cooked meal, only the purchase of the ingredients will count towards GDP. Thirdly, GDP does not account for any change in environmental resources. The implication of this is for instance that extraction of oil and the sale will count positively to GDP, while the deduction of oil reserves will not reduce GDP (Jones, 2014).

2.2 ML algorithms

ML is a subset of artificial intelligence (AI) that provides systems of algorithms and statistical models performing specific tasks. The main advantage of ML algorithms is that they rely on relationships and patterns that they identify, while also having the ability to incorporate a large amount of data. This particular trait makes ML methods highly suitable for learning the underlying, complex structure of data and using this in making predictions for future values. Despite the fact that a lot of these algorithms have been existed since 1970, the field of ML has only gained traction the last decade. This is largely due to that low-cost powerful computational processing and increasing data volumes, have only recently become accessible to a vast majority of people (Economist, 2015). The chosen ML algorithms of BART, GLMNET, GBM and XGBoost have yielded promising results in a wide range of fields such as the health care industry, retail and financial industry. It is therefore interesting to employ these algorithms on GDP-forecasting.

Regression trees

The BART, GBM and XGBoost algorithms build on regression trees. In the core, regression trees are nested if-else conditions, and may be considered as decision trees. The method was first proposed by Morgan and Sonquist (1963) and involves binary recursive partitioning, which is a process that splits the data into partitions and branches and then continue to split each partition into smaller groups. The splits are created by minimizing the sum of squared deviations from the mean in partitions. This process is continued until each node reaches a specified node size, which becomes the terminal node. Note that if the squared deviations from the mean is zero, then that node becomes the terminal node despite the minimum size may not be reached.

2.2.1 BART

BART is a sum-of-trees ensemble Bayesian approach to non-parametric function estimation (Kapelner and Bleich, 2013). In order to approximate a function f, the algorithm uses regression trees to rely on recursive binary partitioning of predictor space into a set of hyperreactangles. The dimension of the predictor space is the same as the number of p variables. While tree-based regression models have an ability to capture interactions and non-linearities, models consisting of sums of regression trees have an even greater ability to capture in interactions and non-linearities. In addition, sums of regression trees have the ability to factor in additive effects in f. The BART model can be written as:

$$Y = f(x) + \epsilon \approx T_1^M(X) + T_2^M(X) + ... + T_m^M(X) + \epsilon, \;\; \epsilon \sim N_n(0, \sigma^2 I_n) \tag{2.2}$$

where Y is the $(n \times 1)$ of response variable, X is the $(n \times p)$ matrix of joined predictor columns and ϵ is the $(n \times 1)$ vector of error element. m denotes the distinct regression trees, T denotes the tree structure and the terminal node is denoted by M. Together T^M represents an entire tree containing the structure and the set of leaf parameters.

2.2.2 GLMNET

In GLMNET each parameter is optimized by minimizing an objective function (Friedman et al., 2010). By using a cyclical coordinate descent the algorithm iterates until convergence (Hastie and Qian, 2014). Let Y_f be the value to forecast, x_i be the matrix of input variables and $x_f = \{x_{f1}, ..., x_{fk}\}^T$ with k number of descriptors. A linear model model for each predicted value can then be written as:

$$E(\beta) = \sum_{f=1}^{n} (y_f - x_f^T \beta)^2 \tag{2.3}$$

The minimizing coefficients are defined by the ordinary least squares method. Due to the case of singularity when $k > n$, regularized regression is applied. The loss function for a GLMNET is defined as follows:

$$E(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{i=1}^{k} \Big((1-\alpha)\beta_j^2 + \alpha|\beta_j| \Big) \tag{2.4}$$

By minimizing the loss function of GLMNET, the coefficients of β can be obtained. α and λ may be used to adjust the model, where $0 < \alpha < 1$. In the case of $\alpha = 1$ the model corresponds to a Ridge regression, and in the case of $\alpha = 0$ the model corresponds to a Lasso regression. The main objective is to minimize the loss function $E(\beta)$ given the parameters α, λ and β.

2.2.3 GBM

GBM builds a model in a step-wise manner while including a differentiable loss function (Friedman, 2001). In other words, the algorithm optimizes a cost function over a given space by iteratively choosing a function that directs in the negative gradient direction. The empirical risk minimization principle states that it is not possible to know the true distribution of data, but we can analyze an algorithm's performance on a known set of training data and thus examine the empirical risk (Stoyanov et al., 2011). In accordance with this principle, the method tries to find $\hat{F}(x)$ that minimizes the average value of a given loss function. Let H be a set of arbitrary differentiable functions on $\mathbb{R}$, then the GBM model is updated in accordance with the following equations:

$$F_m(x) = F_{m-1}(x) - \gamma_m \sum_{i=1}^{n} \nabla_{F_{m-1}} L\Big(y_i, F_{m-1}(x_i)\Big), \tag{2.5}$$

$$\gamma_m = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L\Big(y_i, F_{m-1}\Big((x_i - \gamma \nabla_{F_{m-1}} L(y_i, F_{m-1}(x_i)\Big)\Big) \tag{2.6}$$

where the derivatives are taken with respect to F_i for $i \in \{1,...,m\}$ and γ_m is noted as the step length.

2.2.4 XGBoost

XGBoost is an improved algorithm based on the GBM framework in terms of gradient boosting, which in addition has the ability to construct boosted tree efficiently and operate in parallel (Chen and Guestrin, 2016). While both the GBM and XGBoost follows the gradient boosting principle, the XGBoost algorithm applies a more strict regularization. The main objective of the algorithm is to optimize parameters given an objective function that contains a loss function and a regularization parameter. The regularization term aims to reduce the likelihood of overfitting by controlling the complexity of constructed trees. The complexity of each tree follows the equation:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{2.7}$$

where T is the number of leaves and ω is the vector scores on leaves (Chen and Guestrin, 2016). The structure score, objective function, of the algorithm is defined as:

$$F = \sum_{j=1}^{T} \left(G_j \omega_j + \frac{1}{2}(H_j + \lambda)\omega_j^2 \right) + \gamma T \tag{2.8}$$

where ω_j are independent from each other and the form $G_j \omega_j + \frac{1}{2}(H_j + \lambda)\omega_j^2$ is quadratic.

2.3 Time series regression methods

Unlike ML algorithms, traditional time series methods rely on assumptions of the underlying data and certain prerequisites. While mentioning all of the prerequsities are out-of-scope, the most important are: stationarity and the sequence of uncorrelated error terms with distribution $\epsilon \sim N(0, \sigma^2)$, also called white noise.

2.3.1 The AR model

An AR model is a representation of a type of a random process, and have generally been applied to time-varying events in economics, including stock market forecasting and macroeconomic forecasting. The AR model specifies that the output variable depends linearly on its previous values and a stochastic term, representing a stochastic difference equation (Hilde and Thorsrud, 2014). The notation AR(p) indicates an autoregressive model of p order, and may be expressed as:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \epsilon_t, \quad \epsilon \sim N(0, \sigma^2) \tag{2.9}$$

where φ_i are the parameters of the model, c is a constant and residual element resembles white noise.

2.3.2 The ARIMA model

An ARIMA model is a combination of an Autoregressive (AR) and a Moving Average model (MA), with the inclusion of differencing ability (I). The general non-seasonal notation ARIMA(p,d,q) indicates a p order of the AR part, d degree of first differencing and q order of the MA part. The full model may be written as:

$$y_t = c + \phi y_{t-1} + ... + \phi_p y_{t-p} + \phi_1 \epsilon_{t-1} + ... + \phi_q \epsilon_{t-p} + \epsilon_t, \quad \epsilon \sim N(0, \sigma^2) \tag{2.10}$$

where y_t is the differenced series, ϕ and ϵ indicates the lagged values and errors. Note that by selecting appropriate values for the parameters, we obtain versions of AR models and MA models (Hilde and Thorsrud, 2014).

2.3.3 The VAR model

The VAR model is a stochastic model used to capture the linear interdepencies among multiple time series. VAR models have become widely used in macroeconomics for purposes from forecasting macroeconomic variables and modeling expectation formation in theoretical macroeconomic models. Every factor in a VAR model is managed symmetrically, implicating that all variables are weighted equally. In the model every variable is a linear equation of its own lags and the lags of the other variables (Hilde and Thorsrud, 2014). VAR models are a multivariate generalization of an AR(p) model, and a model of order p can be written as:

$$y_t = \mu + A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + \epsilon_t \tag{2.11}$$

Where A is a $(K \times K)$ coefficient matrices, μ denotes a $(K \times 1)$ vector of intercept terms and e_t is a $(K \times 1)$ dimension vector error terms which we assume resemble white noise with the following properties:

$$\begin{aligned} E[\epsilon_t] &= 0 \\ E[\epsilon_t] &= \begin{cases} \sum_\epsilon & \text{for t = s} \\ 0 \text{ otherwise} \end{cases} \end{aligned} \tag{2.12}$$

3 Data

To guarantee that the input data for our models is rich and representative of the economies we are studying, we should ideally have a wide range of macroeconomic indicators. ML algorithms in particular usually require a lot of data in order to pick up subtle patterns and trends, and this trait of ML algorithms poses some challenges when it comes to the gathering of domestic macroeconomic indicators. Indicators such as GDP, inflation and unemployment rate are usually regularly recorded, while other indicators are not. In order to cope up with this challenge, we only study economies where sufficient data is available with respect to the number of variables and observations. Consequently, we only include well developed economies in our research. To get a representative view on whether the ML algorithms have the ability to forecast GDP growth, we include economies that have had both steady and volatile GDP growth. For instance, Great Britain was included to evaluate whether the models could factor in the recent effect Brexit has had on the nation's GDP growth. USA was included to evaluate whether the models could factor in the Financial Crisis that occured in 2008. Furthermore, we also include the Euro Area, to see whether the inclusion of an economy consisting of a number of economies could improve the forecast performance. In sum, 10 economies will be evaluated: *Australia, Canada, Euro Area, Germany, Spain, France, Japan, Sweden, Great Britain* and *USA*.

For our research, we use macroeconomic data provided by the Quandl database, which is a marketplace for financial, economic and alternative data. The database includes national accounts, monetary, trade and labor statistics, fiscal data and balance of payment accounts. In addition, we have added survey data including Purchasing Manager Indexes (PMI), Business and Consumer Confidence Indexes (CCI) as well as financial market data retrieved from Bloomberg for each country. We also use several other macroeconomic indicators that correlate with GDP growth, among them; employment rate, disposable personal income and new job vacancies. A detailed list of all the variables obtained and the respective sources may be found here (external link), as we found it more convenient to create a webpage rather than include all the variables in appendix.

Country	Observations	Variables	Start	End
Australia	119	62	1990 Q1	2019 Q3
Canada	139	62	1985 Q1	2019 Q3
Germany	99	58	1995 Q1	2019 Q1
Spain	99	59	1995 Q1	2019 Q1
Euro Area	83	53	1999 Q1	2019 Q1
France	99	62	1995 Q1	2019 Q1
Japan	139	61	1985 Q1	2019 Q3
Sweden	119	43	1985 Q1	2019 Q3
Great Britain	139	60	1985 Q1	2019 Q3
USA	199	71	1985 Q1	2019 Q3

Table 3.1: Time series overview

Table 3.1 gives an overview of the number of variables and starting date for each country. In order to fully leverage our forecasting models, we sought to obtain time series for as long as possible while also retaining a sufficient number of variables and observations. As a number of macroeconomic indicators have been recorded with different frequencies and at different times, including the full length of time series would result in an unmanageable number of missing values. For each country we removed variables that included GDP in any version, among them: GDP per capita, GDP noted in American dollar and current account to GDP. Furthermore, we removed variables that contained more than 90% missing values. To fully sought the information in each datapoint, the missing values were imputed using a correlated random forest method, which has the ability to impute missing values given complex interactions and nonlinear relationships between variables. While we are aware of the potential drawbacks of imputation, all of the missing values are only existing in the beginning in each country time series and the chosen imputation method maximizes the likelihood for a given imputed value.

Software and hardware

Data preparation, data handling and performance analysis is performed in **R**, a programming language for general-purpose programming. Our ML algorithms are developed with the package *Caret*. For TS regression models, we use the package *Forecast*. The ML algorithms are trained on R Studio Cloud platform to cope up with the computational power required to run these algorithms, while the traditional TS regression models are run using Central Processing Units (CPU).

4 Methodology

Throughout the chapter we will explain the methodology behind the forecasting models, how we will assess performance evaluation and how we will determine variable importance for the chosen ML algorithms.

4.1 Data split

When building forecasting models it is common practise to separate the data into two portions, namely training data and test data. For forecast purposes it is important that the model is trained on separate observations, the training set, before it is tested on observations the model has not yet seen, the test set. As the test set is held out, the idea is that it should give a reliable indication of whether the model has the ability to forecast unseen data. For our research we are going to assess the forecast performance using a rolling forecasting origin and produce one-step ahead forecast, a method proposed by Tashman (2000). The method is an evaluating technique to which the forecast origin is updated iteratively and the forecast is produced from each origin, as illustrated in figure 4.1. The implication of using this technique is that a given model is reestimated successively for each quarter. We allocate first 75% of the data as training data and then forecast one quarter ahead, while updating the model for each quarter. Using this approach, we will investigate whether the methods have the ability to learn successively for each period.

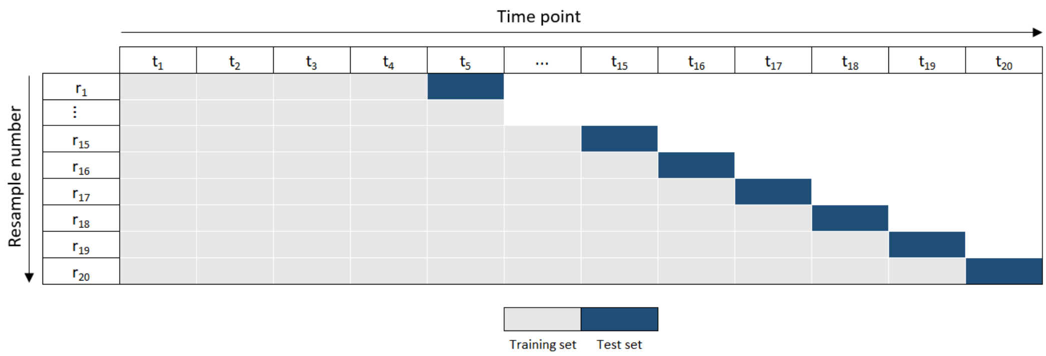


Figure 4.1: Graphical illustration of rolling origin

4.2 Hyperparameter optimization of ML algorithms

For ML purposes hyperparameter optimization is crucial for building adequate forecast models. A hyperparameter may be understood as a parameter where its value is used to control the learning process. A ML algorithm may require different constraints, weights or learning rates to generalize specific data patterns. Furthermore, different ML algorithms have different parameters that need to be set. However, the most challenging part with the parameters is that an analytical framework to calculate appropriate values usually does not exist and thus the choice of these are often subject to trial and error. Hyperparameter optimization finds the appropriate set of a hyperparameters model that minimizes a given loss function and thus yields the optimal model. While many approaches for hyperparameter optimization do exist, we will employ an exhaustive search through a specified subset of parameters for each algorithm. The selection of the different subset of parameters for each algorithm will be drawn from best practice from the literature. For our research we aim to minimize the loss function of RMSE for each ML algorithm.

4.2.1 Parameter tuning - BART

The BART algorithm has five parameters that need to be determined, and these parameters are: number of trees, prior boundary, base terminal node hyperparameter, power terminal node hyperparameter and degrees of freedom.

Number of trees specifies how many decisions trees that should be established. The BART algorithm employs a backfitting algorithm over and through the number of trees. Chipman et al. (2010) found that setting the number of trees equal to 200 usually provides good performance and proposed also investigating two other choices near 200. Their research indicated that increasing the number of trees by 100 drastically improved the predictive performance until at a point where the predictive power slowly started to degrade. Consequently, for prediction purposes the number of trees should not be set too small. We will therefore perform a grid search containing the number of trees equal to 200, 300 and 400.

The prior boundary parameter determines the prior probability that $E(Y|X)$ is contained in the interval (y_{min}, y_{max}) based on the normal distribution. Larger value of k normally results in more shrinkage and thereby a more conservative fit. Chipman et al. (2010)

recommend to set parameter k equal to 2, but also investigating k equal to 1 and 3.

The prior boundary and base terminal node have the ability of shallowing tree structures, such that the complexity of any single tree is reduced (Kapelner and Bleich, 2013). In other words, these parameters may be understood as regularization parameters such that the likelihood for overfitting is reduced. These parameters are incorporated as prior probability $\alpha(1+d)-\beta$ where $\alpha \in (0,1)$ and $\beta[0,\infty]$ for nodes at depth d. The node depth may be understood as the distance from the tree root. Chipman et al. (2010) recommends setting $\alpha = 0.95$ and $\beta = 2$. The grid search of our model will thus contain α value of 0.95, and for the sake of completeness, we will allow the β values range from integer 1 to 4.

Lastly, degrees of freedom are for the inverse $\tilde{\chi}^2$ prior. On this point Chipman et al. (2010) recommended against setting $v < 3$, and showed that lower values of v led to overfitting. In their research, they showed that, generally best predictive performance was achieved using either $v = 3$ or $v = 10$, and therefore we will vary v from the integers 3 to 10.

4.2.2 Parameter tuning - GLMNET

For the GLMNET algorithm, we need to determine two parameters, namely the mixing percentage and the regularization parameter.

The mixing percentage ranges from 0 to 1, and specifies what weight to give the two types of regression, respectively Ridge and Lasso regression. Setting this parameter equal to 0 will yield a Ridge regression, while a parameter value of 1 will yield a Lasso regression. As described in the 2.2.2, the main advantage of a GLMNET model is its ability to draw aspects from both the Lasso and the Ridge regression, shrinking the coefficients and setting the coefficients' of less contributing variables equal to zero. Consequently, we apply a grid search making the mixing percentage range from 0 to 1, with the interval of 0.1.

The regularization parameter, λ, represents the penalty term. As the GLMNET model combines both the Ridge and Lasso model, the penalty term of both of these models are included. For the Ridge regression this represents the term that shrinks those variables that have minor contribution to the outcome, while for the Lasso variable this represents the shrinkage term that reduces those coefficients that have minor contribution to the outcome to be equal to 0. To test a sufficient amount of λ values, we construct a grid search ranging from $[10^{-3}, 10^{3}]$ with the length of 1000, meaning that every combination in between is constructed.

4.2.3 Parameter tuning - GBM

For the GBM algorithm we have to determine four parameters. These parameters include number of boosting iterations, maximum tree depth, shrinkage and minimum terminal node size. The main goal is finding a model that is sufficiently complex, while also minimizing the chance of overfitting the data by employing a loss function of RMSE with cross validation.

The number of boosting iterations represents the number of trees in the additive model of GBM. Increasing the number of iterations, will increase the representational ability on the given training set. Generally, increasing boosting iterations reduces the training error, while setting it too high may result in overfit (Natekin and Knoll, 2013). The goal is to find the optimal number of trees that minimizes the loss function, RMSE in our case, with cross validation. To test a sufficient amount of models, we apply a grid containing number of trees from 50 to 1 500, with the interval of 50.

The maximum tree depth may be understood as the maximum number of terminal nodes. With increased number of terminal nodes, comes increased ability of representing complex functions (Natekin and Knoll, 2013). The more terminal nodes a tree has, the fewer observations tend to be in a region. This further generally leads to a higher variance and increased complexity which yields overfit. On the other hand, if the number of terminal nodes are too few, the tree might not be able to capture sufficient orders of interactions leading to bias. Consequently, there is a trade-off between variance and bias, and the goal is to make the model not overfit the training sample. In the literature, researchers usually have set this parameter to 5 (Natekin and Knoll, 2013). Thus, we will investigate models ranging from 1 to 5.

The shrinkage parameter represents the learning rate, and controls how quickly the algorithm proceeds down the gradient descent. Generally, the parameter shrinks the added bias at each iteration. By setting the parameter too high, we risk the algorithm to learn too much of the structure on the early iterations leading to a high variance. Thus, by lowering the learning rate, the algorithm becomes able to add a number of trees to the additive tree before overfitting the data (Natekin and Knoll, 2013). We will investigate shrinkage penalties from 0 to 1, with an interval of 0.1.

Lastly, the minimum terminal node size specifies the minimum number of observations in a terminal node. In a number of software packages, the default value for the minimum terminal node size is 1 for classification predictions and 5 for regression predictions. Natekin and Knoll (2013) emphasized that these values provide good predictions results, but that the performance could be improved by tuning it. Normally, one would set a higher terminal node to reduce the computational power required. Therefore, we will will investigate terminal node equal to 1 as we are running the algorithms on the R Studio Cloud platform.

4.2.4 Parameter tuning - XGBoost

XGBoost is an extensive ML algorithm that requires a number of parameters to be tuned. These parameters include number of boosting iterations, max tree depth, shrinkage, minimum loss reduction, subsample percentage, subsample ratio of columns, fraction of trees dropped and probability of skipping drop-out and minimum sum of instance weight.

The algorithm do have some of the parameters as the GBM model we developed in 4.2.3. These parameters include number of boosting iterations, max tree depth and the shrinkage rate. Higher values of these parameters generally increase the representational ability and learning ability, while also risking the model to overfit the training sample leading to poor test performance. For the arguments presented in 4.2.3, we apply the same values for these parameters in the grid search. As presented, for the number of boosting iterations we will investigate iterations between 50 and 1500 with an interval of 50, max tree depth will vary between the integers 1 to 5 and the shrinkage rate will vary between 0.1 and 1 with an interval of 0.1.

The minimum loss reduction is a pseudo-regularization hyperparameter and controls the complexity of a tree. This parameter specifies the minimum loss reduction to further partition a leaf node of a tree. When this parameter is applied, the algorithm will build a tree to the max depth specified, but then prune the tree and remove those splits that do not meet the specified value. This parameter ranges from 0 to ∞, where 0 represents no regularization. Greenwell (2019) recommends exploring values from 0 to 20, and we follow this recommendation by exploring values from 0 to 20 with an interval of 5.

The subsample percentage and subsample ratio of columns specifies the subsample ratio of the training data and subsample ratio of columns when constructing a tree. These values range from 0 to 1, and aggressive subsampling in terms of higher values has shown to lead to good performance. Generally, the performance depends on the multicollinearity of the predictors in the training data. If there are fewer relevant predictors for a given outcome, higher values of subsample normally yield better performance as the parameter makes the algorithm more likely to choose those features with the strongest signals (Greenwell, 2019). The recommended values in this case ranges from 0.5 – 0.8, and we will therefore apply these values in our grid search.

The fraction of trees dropped and probability of drop-out is an alternative approach to reduce overfitting, and may also be understood as regularization parameters. When the algorithm runs the boosting iterations, it typically adds considerably weight to the first constructed trees, while trees added later gains considerable less weight. This usually leads to overfitting and the idea with dropout is to build an ensemble by randomly dropping trees in the boosting sequence (Srivastava et al., 1970). Both the fraction of trees dropped and probability of drop-out range from 0 to 1, and we will investigate the values in between as there are not any guidelines in the literature on what these values that normally yields good performance.

The minimum sum of instance weight specifies when the building process of model should give up further partitioning. If a tree partitioning results in a leaf node with the sum instance weight less than the given parameter, then this lead node does not add any value to the model. In other words, this represents the number of instances that is needed in every leaf in each tree. Consequently, higher values for this parameter yield a more conservative model. This parameter ranges from 0 to ∞. To reduce the complexity, we will only apply the default value of 1 to this parameter in the grid search.

4.3 Framework for time series regression models

4.3.1 The AR model

To estimate an AR(p) model, we need to estimate the optimal number of lags to include in the autoregression. Including too few lags may result in omitting valuable information leading to the residuals easily becoming autocorrelated. Consequently, everything not included as an independent regressor will end up in the residual (Hilde and Thorsrud, 2014). On the other hand, if we include too many lags we end up estimating more coefficients than needed, which in turn introduces other estimation errors. For the AR model we will use statistical testing procedures. We will start with specifying a model with maximum number of lags, and then investigate whether the coefficient on the last lag is significantly different from zero using a t-test. If the last lag is not significantly different from zero we will exclude it and estimate an AR $(p-1)$ model and continue in this fashion for $p = [P, ..., 1]$.

4.3.2 The ARIMA model

To estimate an ARIMA model we apply a method that is inspired by the Hyndman algorithm (Hyndman, 2008) and the Box-Jenkins method (Din, 2015). Model identification of an ARIMA model consists of deciding the AR and MA parts of the model. Number of differencing (d) will be determined by evaluating the non-stationarity of the data, which will be evaluated using a KPSS-test. The autoregressive (d) and moving average (q) terms are investigated using an Autocorrelation function (ACF) and a Partial Autocorrelation function (PACF). To determine which model that will be utilized as part of this study, we apply the Aikakes information criteria (AIC), as this criterion suggests the optimal model number of lags and parameters to be assessed in the models. AIC is based on information theory where a statistical representation of a given process will almost never be exact, and therefore AIC estimates the relative amount of information lost by a given model. The less information lost, the higher quality of a given model is attained. AIC can be derived by:

$$AIC(p) = ln\left(\frac{SSR(p)}{T}\right) + (p+1)\frac{2}{T} \tag{4.1}$$

where $SSR(p) = \sum_{t=1}^{T} \hat{\epsilon}_t \hat{\epsilon}_t$. Diagnostics will then be performed to evaluate suitability of the fitted model, namely investigating whether the residuals resemble the properties of white noise using Ljung-box test. If the residuals do not resemble white noise, we will start the algorithm from the beginning and search for other models.

4.3.3 The VAR model

In model identification of a VAR model we first have to specify which variables to include in the model. VAR models may easily become heavily parameterized ending up with too many parameters to estimate relative to the observations in the data, resulting in the degrees of freedom problem (Bernshtein, 1967). Generally, including more than 6 variables in a VAR model may result in largely imprecisely estimated parameters (Hilde and Thorsrud, 2014). As such, VAR models are not able to include all variables of potential interest and we will draw existing literature when choosing variables. Firstly, similar to Marcellino et al. (2006) we will build a VAR model that includes GDP, unemployment rate and inflation where we will transform the last two variables to represent quarterly percentage changes. Marcellino et al. (2006) also indicated that including more variables did not yield better results, as simple models in general are often marginally less precise than complex models. Research done by Andersson (2007) also indicated that unemployment rate and inflation were the best predictors for GDP growth. Secondly, introduced by Christiano et al. (1999) when investigating the effects of monetary policy shocks, we will build a VAR model containing GDP growth, inflation rate and interest rate. This relationship stems from standard economic theory, that contractionary monetary policy shocks that increases (decreases) interest rate will in turn have a temporary negative (positive) effect on GDP and inflation.

Furthermore, the appropriate lag length has to be determined. Common practices include choosing a large lag length a priori and investigate robustness to results by reestimating with shorter lags and using Hannan-Quinn Information Criteria (HQ). HQ is given by:

$$HQ = -2L_{max} + 2kln\Big(\ln(n)\Big) \tag{4.2}$$

Where L_{max} is the log-likelihood, k is the number of parameters and n is the number of observations. When the optimal number of lags has been resolved, the parameters of the VAR model should be assessed. The most common method is using an Ordinary Least Square Estimator (OLS), as it is the natural estimator (Hilde and Thorsrud, 2014). A large lag length relative to the number of observations will lead to poor and inefficient estimates of the parameters. Contrarily, a lag that is too short will result in misspecified parameters and biased OLS estimates, resulting in spurious significance of the parameters (Canova, 1995). Consequently, for our research we will determine optimal number of lags using HQ and assess the parameters using OLS. Whether to include insignificant lags will be assessed according to each particular fitted model, and we will also examine the evidence of autocorrelation in the residuals using Ljung-box-test.

4.4 Forecast performance

To assess the forecast performance we use three forecast evaluation metrics. The first is the Root Mean Squared Error (RMSE), which calculates the square of average differences between predicted and actual observations. RMSE can be calculated by:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_t - A_t)^2} \tag{4.3}$$

Where A_t is the actual value at time t, F_t is the forecasted value at time t and N represents the number of forecast points. Furthermore, we also use the Mean Absolute Percentage Error (MAPE). MAPE is calculated by:

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{A_t - F_t}{A_t}\right| \tag{4.4}$$

Moving on we also use the Mean Directional Accuracy (MDA), which compares the forecast direction in terms of upwards or downwards to the actual direction. This metric may for instance be used by a monetary authority to determine whether to raise or lower the interest rate given that inflation is expected to rise or drop. MDA is calculated by:

$$MDA = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{sign(A_t - A_{t-1}) == sign(F_t - A_{t-1})} \tag{4.5}$$

where sign is the sign function and $\mathbb{1}$ is the indicator function. In terms of GDP forecasting, RMSE is particularly useful as it gives relatively high weight to undesirable large errors. MAPE is included because it gives an intuitive interpretation in terms of relative error, as it is noted in percentage. While both RMSE and MAPE provides information about the accuracy and the value of the forecasts, it may be equally crucial to accurately forecast the direction of change and thereby MDA is also included. Using these three metrics in combination with plots to assess the forecast performance, will give a more reliable assessment of how the models are performing. We will however put most emphasis on RMSE, MDA and the plots.

4.5 Variable importance

When assessing the performance of each ML-algorithm, it is also interesting to see which variables that are contributing the most for each ML model. While we cannot draw causal inference by assessing the significance of individual variables by employing standard statistical tests (Chakraborty and Joseph, 2017), we can assess which variables that drives the predictions of the chosen algorithms. The following methods are used to estimate the contribution of variables in each algorithm:

- **GLMNET**: For linear models such as GLMNET, the absolute value of the t-statistic for each parameter is used.
- **BART**: The reduction in the given RMSE loss function attributed to each variable is tabulated and then the sum is returned. Since there might be some variables that are not used in a split, but still may be considered as important, the top competing variables will also be tabulated and split.
- **GBM and XGBoost**: For both of these algorithms, we apply a permutation test. The method randomly permutes each variable at a time and computes the associated reduction in forecast performance. This is then repeated for each boosting iteration, and then the difference is both averaged over all trees and normalized by the standard error.

Note that we will only apply variable importance when we examine the results of the forecasts produced, and to assess the ML algorithm that yields best forecast performance.

5 Analysis and Results

In the analysis, we will first examine the overall results. For this part we will focus on the performance of each model on an average level rather than for each particular country. Thereafter, we will do a more comprehensive visual examination of selected economies by studying the plots of the forecasts in combination with the forecasting accuracy. The plots of rest of the economies may be found in appendix. We will also do an assessment of the top three variables that drives the best performing ML algorithm.

5.1 Overview of forecast accuracy

From the table 5.1 we see that the traditional TS regression models in general yield lowest RMSE. The VAR models yield RMSE averages of respectively 0.563 and 0.547, which makes them the best performing models. This particular finding is interesting due to two reasons. Firstly, all of the ML algorithms are fed with a number of macro economic variables. Despite the access to variables such as surveys responses from industry leader and consumers regarding the state of the economy, the algorithms are not able to outperform the VAR models. However, the algorithms are able to outperform the AR and ARIMA models. This results is not surprising as these models represents the naïve models for our research. Moreover, of the ML algorithms, the XGBoost algorithm has the lowest RMSE on average, but it represents an increase in RMSE of respectively 10% and 13% compared to VAR_1 and VAR_2. In terms of RMSE, the traditional TS regression models are on average yielding lower RMSE than ML algorithms indicating their suitability of forecasting GDP growth.

Similar to the forecasting results of RMSE, we observe that the traditional TS regression models are also yielding better results with regards to the forecasting metric of MAPE. Interestingly, the relatively less sophisticated model of ARIMA are outperforming the other models in 4/10 cases, while the VAR_1 on average yields lowest MAPE. We further notice that while some of the ML algorithms are able to outperform the TS regression models, the clear majority of TS regression models outperform the ML algorithms. The results further indicate that traditional TS regression models are yielding forecasting errors than ML-algorithms and are better suited to forecast GDP growth.

		TS Regression Models				ML Algorithms			
		AR	ARIMA	VAR_1*	VAR_2**	BART	GBM	GLMNET	XGBoost
RMSE	Australia	1.184	0.631	0.823	0.776	0.756	0.898	0.780	0.757
	Canada	0.706	0.519	0.646	0.521	0.800	1.309	0.801	0.853
	Germany	0.784	0.622	0.619	0.647	0.566	0.562	0.536	0.509
	Spain	0.118	0.169	0.088	0.093	0.224	0.212	0.191	0.183
	Euro Area	0.175	0.104	0.099	0.090	0.185	0.186	0.178	0.164
	France	0.820	0.749	0.384	0.334	0.764	0.780	0.710	0.730
	Japan	0.987	0.881	0.853	0.840	0.930	0.903	0.884	0.902
	Sweden	0.609	0.528	0.386	0.427	0.551	0.587	0.541	0.516
	Great Britain	1.454	1.056	1.073	1.078	1.170	1.167	1.141	1.116
	USA	0.611	0.578	0.659	0.665	0.509	0.475	0.495	0.469
	Column average	0.745	0.584	0.563	0.547	0.646	0.708	0.626	0.620
MAPE	Australia	8.21%	1.20%	6.28%	6.56%	1.47%	2.31%	3.09%	3.11%
	Canada	2.22%	1.40%	2.00%	1.41%	0.94%	1.29%	1.61%	1.24%
	Germany	1.15%	1.13%	1.38%	1.90%	1.35%	1.35%	1.36%	1.34%
	Spain	0.22%	0.36%	0.19%	0.19%	0.30%	0.33%	0.41%	0.37%
	Euro Area	0.30%	0.16%	0.24%	0.25%	0.36%	0.40%	0.37%	0.32%
	France	2.97%	9.67%	6.24%	12.82%	12.91%	7.63%	10.10%	17.59%
	Japan	6.13%	1.07%	2.99%	1.68%	1.63%	1.94%	1.33%	1.74%
	Sweden	27.31%	24.27%	14.77%	21.73%	31.36%	23.35%	26.21%	14.92%
	Great Britain	4.04%	2.78%	3.51%	3.39%	2.59%	3.04%	2.36%	2.24%
	USA	1.10%	1.11%	1.36%	1.06%	0.89%	0.82%	0.98%	0.86%
	Column average	5.36%	4.31%	3.90%	5.10%	5.38%	4.25%	4.78%	4.37%
MDA	Australia	48%	79%	45%	97%	66%	59%	52%	69%
	Canada	82%	91%	79%	88%	71%	65%	53%	62%
	Germany	54%	50%	25%	33%	54%	54%	54%	67%
	Spain	67%	46%	67%	79%	75%	75%	63%	71%
	Euro Area	55%	70%	65%	70%	70%	55%	50%	80%
	France	50%	67%	88%	83%	58%	63%	63%	50%
	Japan	50%	71%	56%	83%	53%	53%	68%	44%
	Sweden	45%	31%	66%	72%	53%	53%	68%	44%
	Great Britain	47%	76%	62%	88%	53%	50%	50%	68%
	USA	63%	61%	51%	90%	55%	61%	53%	69%
	Column average	56%	64%	60%	78%	61%	59%	57%	62%

* VAR_1 is a multivariate model containing GDP growth, growth rate of inflation and growth rate of unemployment rate
** VAR_2 is a multivariate model containing GDP growth, growth rate of inflation and interest rate
Note: The grey shadow indicates the best performing model by row

Table 5.1: Forecast accuracy

For most of the cases the MDA metric is over 50%, which is better than a random guess. The VAR_2 has the highest MDA in 6/10 cases, which makes it the best average performer in anticipating the direction of GDP growth. The implication is that the interest rate and unemployment rate seem to be adequate indicators on whether one should expect a rise or decline in GDP growth for the selected economies. Furthermore, except from the XGBoost algorithm, none of the algorithms are able to outperform the performance of the traditional TS regression models on country-level. The forecasting results in general indicates that building complex models in GDP growth forecasting, do not necessarily yield better forecast performance. In cases, the ML algorithms perform better than the traditional TS regression models, the results are often only marginally better.

When examining the performance of each model on each country, we notice some certain aspects. Compared to the results of ML algorithms, the TS regression models yield better forecast accuracy for Canada. Furthermore, both the TS regression models and ML algorithms yield seemingly precise forecast accuracy for Spain and Euro Area. Neither of the models seem to yield any adequate results for Great Britain. Lastly, we also note that the forecast performance of ML algorithms for USA is marginally better than the forecast performance of TS regression models. We will therefore further examine these economies by visual examination, to investigate whether there are some traits that can explain the performance of the models. Furthermore, we will also investigate whether there are certain aspects that either the TS techniques or ML algorithms are better equipped to address. The plots for rest of the economies may be found in Appendix

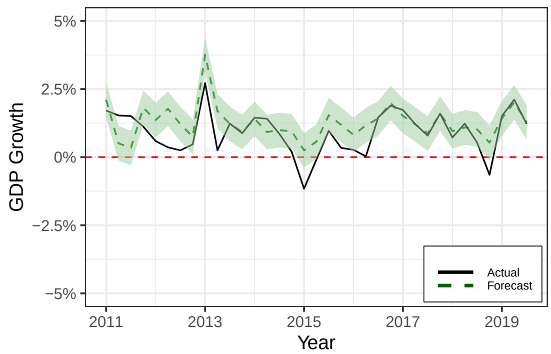
5.1.1 Our findings relative to previous literature

Our research have some overlap with Jung et al. (2018) on economies studied. These economies include Germany, Great Britain, Spain and USA. Except from the results on the Spain time series, the forecast results of our best performing ML algorithm represents an increase in terms of RMSE compared to the results found by Jung et al. (2018). This particular difference is interesting due to two reasons. Firstly, we utilize more data for each time series than Jung et al. (2018), meaning that the algorithms should have more data to learn patterns and relationships from. Secondly, while they do employ other ML algorithms, their results show that ML algorithms are superior in forecasting GDP compared to TS regression models. It is difficult to determine why our research indicate otherwise than Jung et al. (2018) as they do not elaborate on exactly which variables they have utilized, and the possible reasons might be due to differences in hyperparameter optimization or quality of variables included.

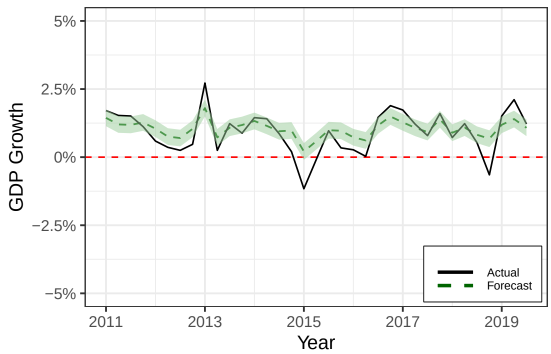
5.2 Examination of selected economies

5.2.1 Canada

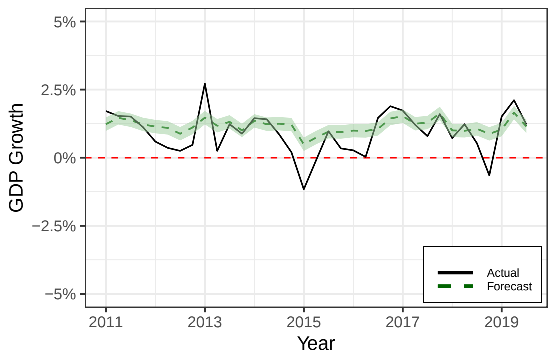
As we can see from Figure 5.1 all of the TS regression models yield adequate forecasts. Except from the VAR_1 model, all of the models successfully forecast the sharp upturn in the Canadian economy in 2013. The AR-model seems to forecast the sharp upturn a bit too optimistic, while from 2016 on the forecasts are seemingly accurate. The VAR_1 yield RMSE of 0.521 and a MDA value of 79%. This indicates that interest rate and inflation are highly suitable variables for forecasting GDP growth in a VAR model. Interestingly, all of the TS regression forecasts also pick up the sharp downturn in 2015. While all of the forecasts underestimate the magnitude of these sharp turns, the results indicate that employing these TS regression models on data for Canada may give an indication of which direction the GDP growth is heading.



(a) AR

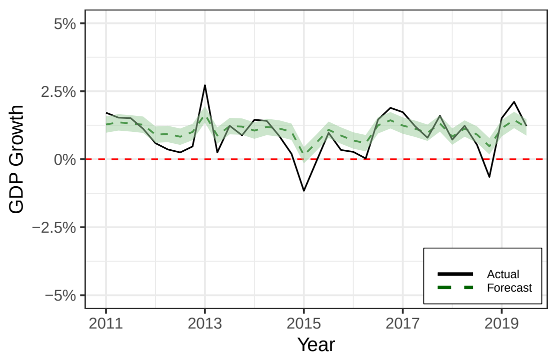


(b) ARIMA



(c) VAR_1*

* VAR_1 is a multivariate model containing GDP growth, growth rate of inflation and growth rate of unemployment rate



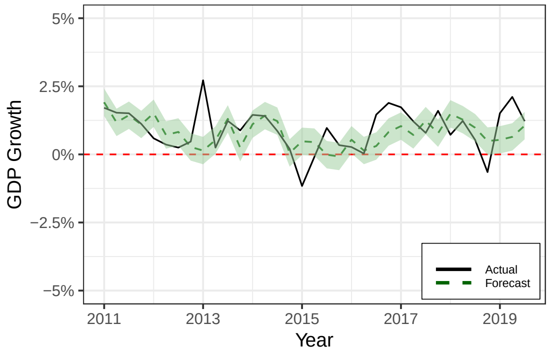
(d) VAR_2**

** VAR_2 is a multivariate model containing GDP growth, growth rate of inflation and interest rate

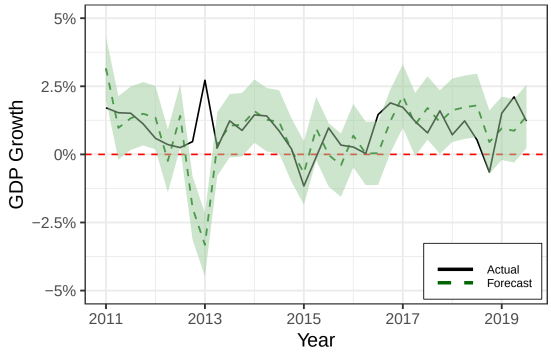
Figure 5.1: TS model forecasts - Canada

Note: The green shadow indicates the area one standard deviation from the forecast point value

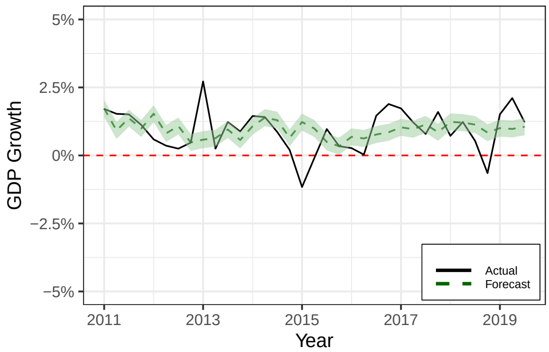
We see from the figure 5.2 that the forecast performance of each ML algorithm is generally lower compared to the performance of TS regression models. For the ML algorithms, the BART and GLMNET are the best performing with respect to RMSE. This is further confirmed in Figure 5.2, where we also notice that the XGBoost is performing at an adequate level. Interestingly, the GBM model forecasts a recession in 2013, which also deteriorates the value of the RMSE. Despite the fact that the model is updated iteratively for each quarter, the model forecasts a sharp declining GDP growth. For the GBM model, the three most important variables are: monthly percent change in consumer prices (inflation rate), monthly total change in retail sales and total number of employment change. These variables make sense in a macroeconomic perspective, as all of these variables either directly or indirectly affect the GDP growth a country. The monthly change in retail sales and total number of employment change, affects how much may be contributed to GDP, while the inflation rate affects the consumption which in turn affect GDP. In the quarters leading up to 2013Q1 Canada experienced shifting inflation rate change, retail sales change and employment change. For some of the quarters, two or less of these variables were negative, which explains why the GBM model forecasted a recession in 2013. Although the GDP growth up until 2013Q1 was positive, the underlying variables indicated otherwise. This shows that despite the poor forecast performance of the ML algorithms, assessing the variable importance of the algorithm may yield some interesting insight.



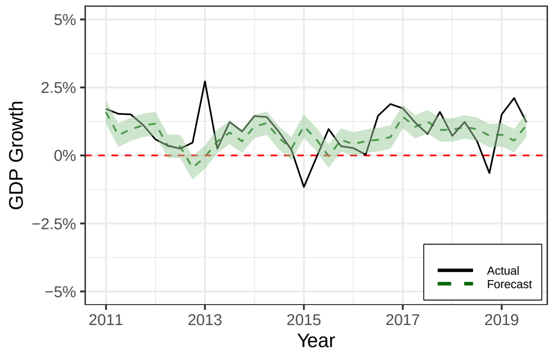
(a) BART



(b) GBM



(c) GLMNET



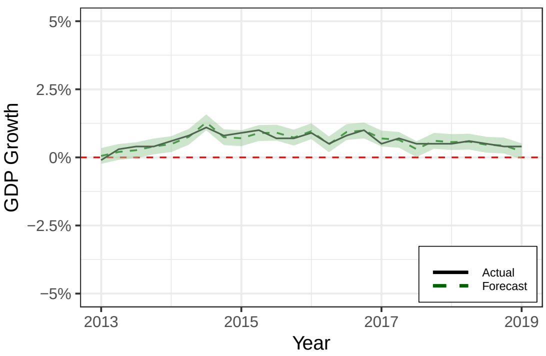
(d) XGBoost

Figure 5.2: ML forecasts - Canada

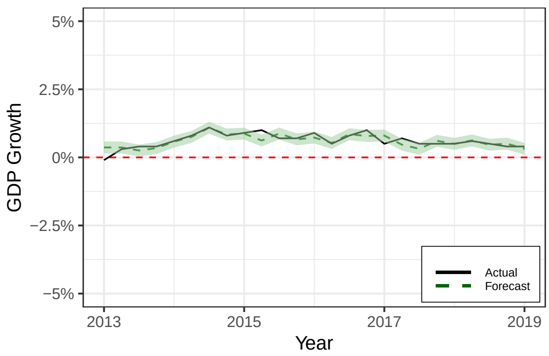
Note: The green shadow indicates the area one standard deviation from the forecast point value

5.2.2 Spain

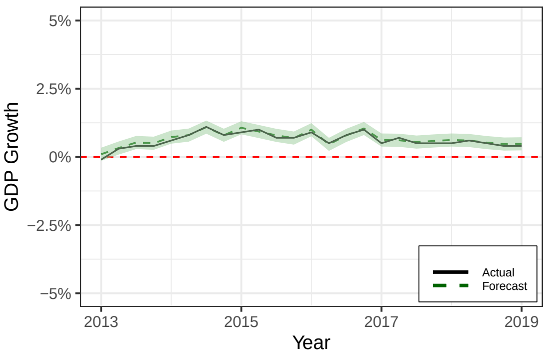
As we can see from Figure 5.3, the GDP growth of Spain has been relatively stable since 2013 and has not been subject to volatility since the Great Recession hit in 2007. While all of the TS forecasts perform well, the VAR models produce the lowest RMSE and highest MDA values. Furthermore, including the AR model and all of the VAR forecast points are one standard deviation away from the true value of GDP growth. This emphasizes that the selected predictors in the VAR models are equipped to forecast GDP growth. Compared to Canada, the RMSE is significantly lower for all of the TS models, while the forecast accuracy of MAPE is somewhat higher, indicating that these models are a better fit for the economy of Spain than for the economy of Canada. This difference in performance seems to indicate that the TS regression models might be better equipped to forecast stable GDP growth rather than fluctuating GDP growth.



(a) AR

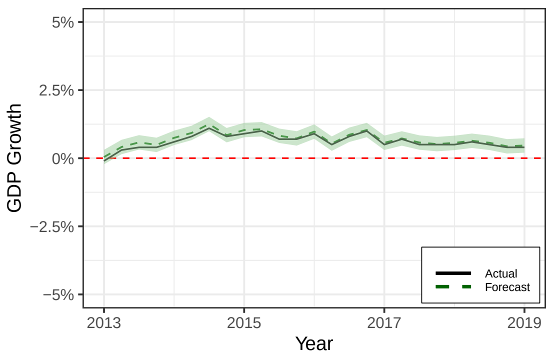


(b) ARIMA



(c) VAR_1*

* VAR_1 is a multivariate model containing GDP growth, growth rate of inflation and growth rate of unemployment rate



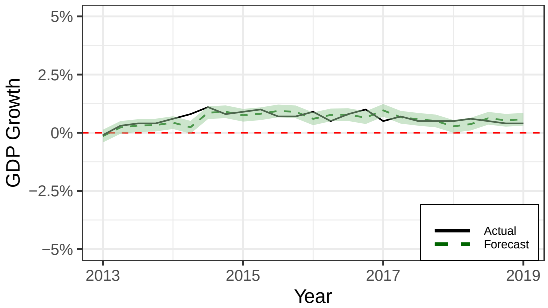
(d) VAR_2**

** VAR_2 is a multivariate model containing GDP growth, growth rate of inflation and interest rate

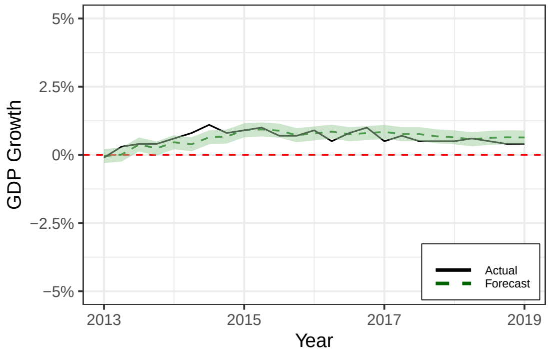
Figure 5.3: TS model forecasts - Spain

Note: The green shadow indicates the area one standard deviation from the forecast point value

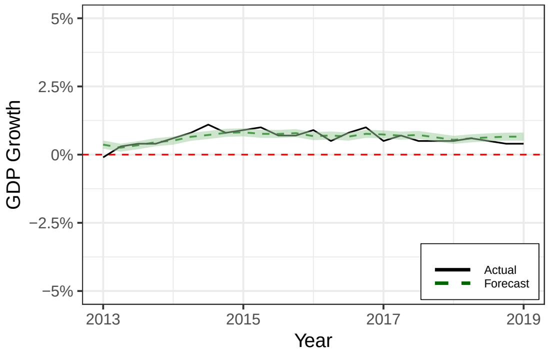
Moreover, the ML algorithms also yield significantly better forecast results for the Spain time series than for the Canadian time series. Compared to the result for the Canadian time series, we see that GBM algorithm produce more accurate forecast points, indicating that the GBM algorithm might be more appropriate forecasting a stable GDP growth. Moreover, none of the forecast points of the ML algorithms are within one standard deviation from the true value of GDP growth. Of the ML algorithms, the XGBoost algorithm produces the best forecast performance with the lowest forecast accuracy. Employing feature selection on the algorithms reveals that the three most important variables for this algorithm is the percent change in total monthly retail sales relative to one year earlier, market value of new orders of manufactured goods in Spain by month and total value of bank loans to the private sector of Spain. Even though one may make the case that all of these variables are important for GDP growth, we cannot avoid the fact that all of the traditional TS regression models outperforms every single ML algorithm. Despite the fact that the ML algorithms all have the ability of regularization and a large number of variables to draw information from, the traditional TS techniques are still outperforming the ML algorithms.



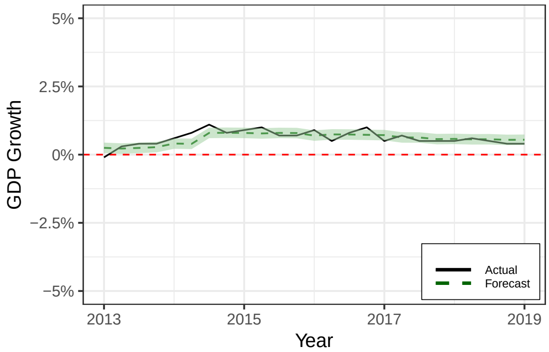
(a) BART



(b) GBM



(c) GLMNET



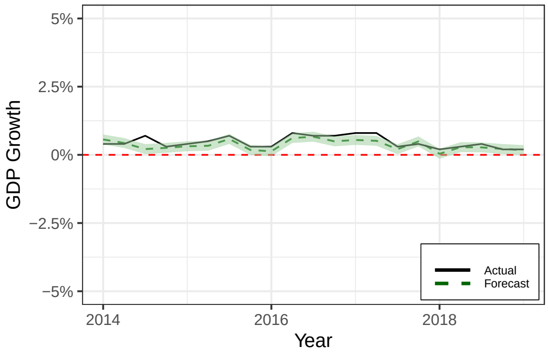
(d) XGBoost

Figure 5.4: ML forecasts - Spain

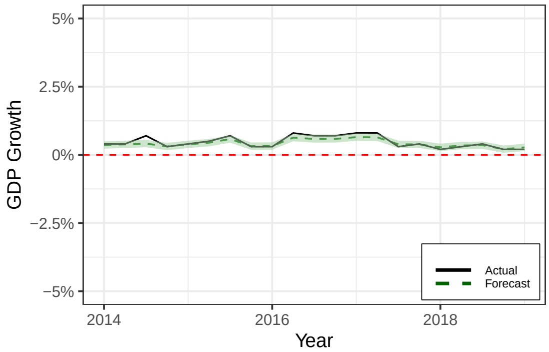
Note: The green shadow indicates the area one standard deviation from the forecast point value

5.2.3 Euro Area

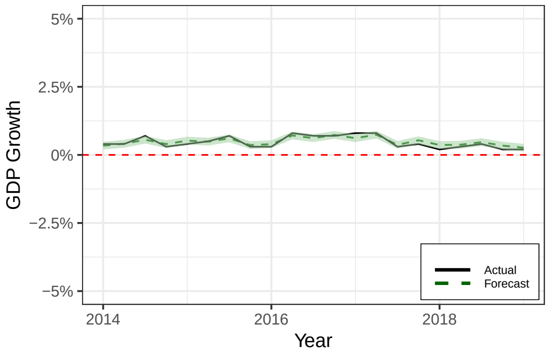
Similar to the GDP growth of Spain, the GDP growth of the Euro Area is relatively stable. This particular trait makes both the TS regression and ML algorithm produce quite accurate forecasts. Both the TS regression models and ML algorithms yield low RMSE, low MAPE and relatively high levels of MDA, indicating low forecast error and high directional accuracy. While all of the TS regression models are performing well, the TS regression model VAR_2 has the best forecast performance. This indicates that the growth rate of inflation and interest rate are well-suited on forecasting the GDP growth of the Euro Area. It should also be mentioned that the VAR_2 is producing marginally better forecasts than the ARIMA model and the VAR_1 model. As we can see in Figure 5.5 all of the models successfully forecast the short-term fluctuations that occur. As a result, these models and their properties should not be disregarded in terms of forecast performance for forecasting the GDP growth of the Euro Area.



(a) AR

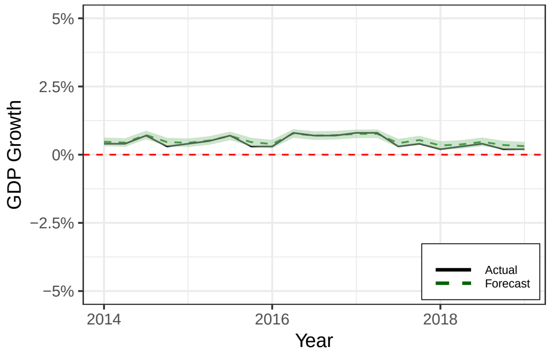


(b) ARIMA



(c) VAR_1*

* VAR_1 is a multivariate model containing GDP growth, growth rate of inflation and growth rate of unemployment rate



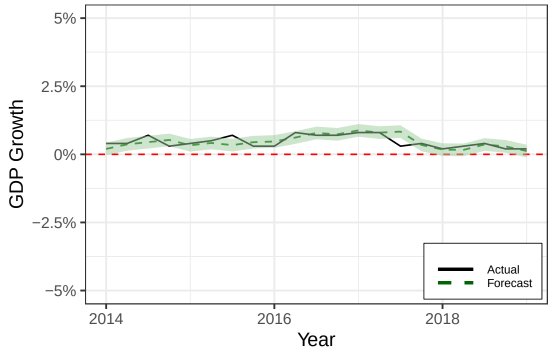
(d) VAR_2**

** VAR_2 is a multivariate model containing GDP growth, growth rate of inflation and interest rate

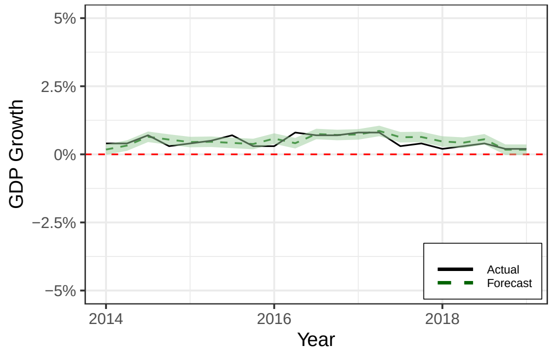
Figure 5.5: TS model forecasts - Euro Area

Note: The green shadow indicates the area one standard deviation from the forecast point value

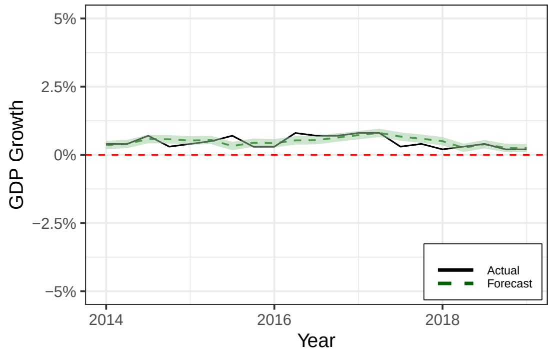
From Figure 5.6 we see that all of the ML algorithm forecasts the GDP growth seemingly well. Of all the models, the XGBoost model is yielding marginally better forecast performance. Examining the three most important variables that drive the forecasts of the XGBoost model yield the following three variables: fraction of eligible workers who have been unemployed and looking for work for an extended period of time, monthly index tracks economists' expectations for the European economy and growth rate of the prices in Euro area. Putting most emphasis on these variables, the XGBoost algorithm is able to outperform the other algorithms. However, the XGBoost algorithm is not able to outperform any of the traditional TS regression models. This is especially surprising for univariate models of AR and ARIMA, as the XGBoost has access to a large number of macroeconomic indicators as explanatory variables. This further underlines that more complex models do not necessarily yield better forecast performance than the simpler ones.



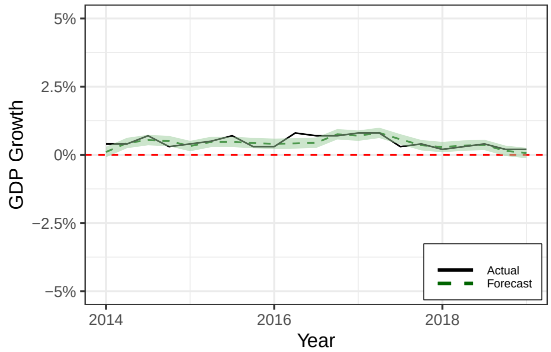
(a) BART



(b) GBM



(c) GLMNET



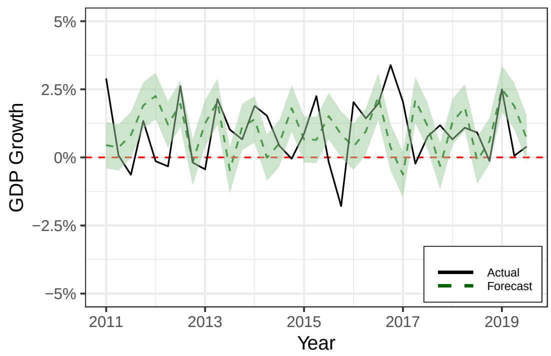
(d) XGBoost

Figure 5.6: ML forecasts - Euro Area

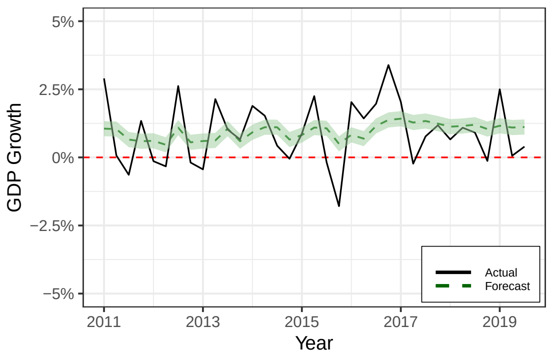
Note: The green shadow indicates the area one standard deviation from the forecast point value

5.2.4 Great Britain

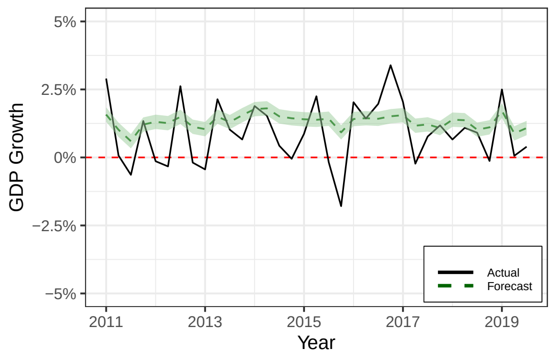
The economy of Great Britain has been volatile in recent years, as can be seen in Figure 5.7 and 5.8. This may explain why both the TS regression models and ML algorithms have difficulties in forecasting the GDP growth of Great Britain. Except from the AR-model, none of the models attempts forecast short-term fluctuations at the same level. Despite the AR-model capturing the fluctuations at a relatively reasonable rate, the RMSE is detoriated, making the AR-model the worst performing model. This further emphasizes the need of examining forecasts through both accuracy metrics and visual inspection. Both the VAR_1 and VAR_2 have smoothed the forecast, indicating that the chosen variables in the VAR-models may not be relevant for forecasting the GDP growth of an economy as volatile as that of Great Britain.



(a) AR

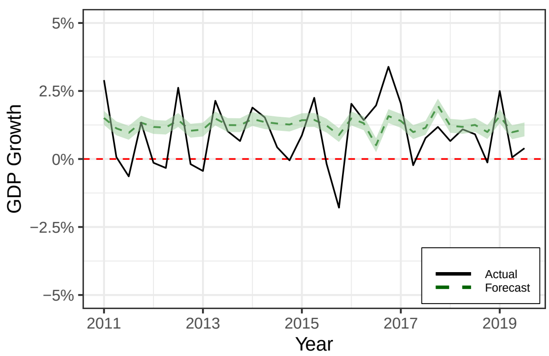


(b) ARIMA



(c) VAR_1*

* VAR_1 is a multivariate model containing GDP growth, growth rate of inflation and growth rate of unemployment rate



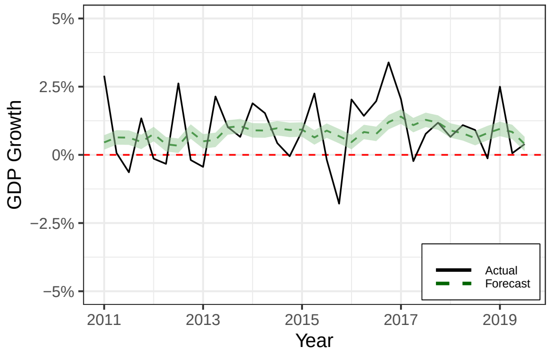
(d) VAR_2**

** VAR_2 is a multivariate model containing GDP growth, growth rate of inflation and interest rate

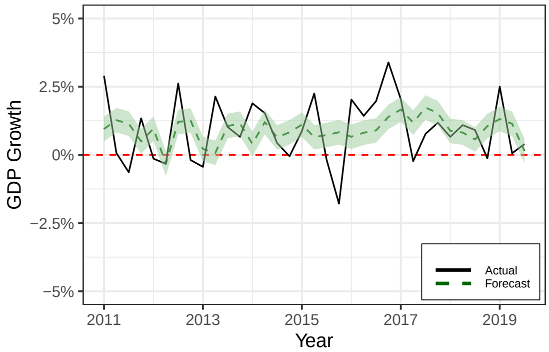
Figure 5.7: TS model forecasts - Great Britain

Note: The green shadow indicates the area one standard deviation from the forecast point value

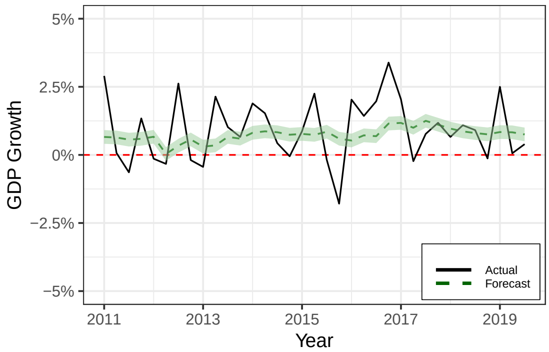
Similar to the TS regression forecasts, the ML forecasts do not capture the short-term fluctuations in the GDP growth of Great Britain. Compared to all of the forecasts, XGBoost is yielding marginally improved forecast accuracy. The feature selection of this model yields that year-over-year percent change in the monthly industrial production, quarterly number of corporate bankruptcies and monthly ratio of export prices to import prices are the top three variables that drives the XGBoost model. While we can make a argument that all of these variables are important for GDP growth, it seems reasonable to assert that there are other variables that may be more accurate in GDP growth forecasting. These findings suggest that there might be other variables that are not included in the Great Britain time series, that are more relevant for forecasting the GDP growth of Great Britain in the given horizon. On the other hand, these forecast errors may also be due to political and social shocks that are difficult to foresee. One aspect is clear, the results indicate that the chosen models and the variables included are relatively inadequate in forecasting the GDP growth of Great Britain.



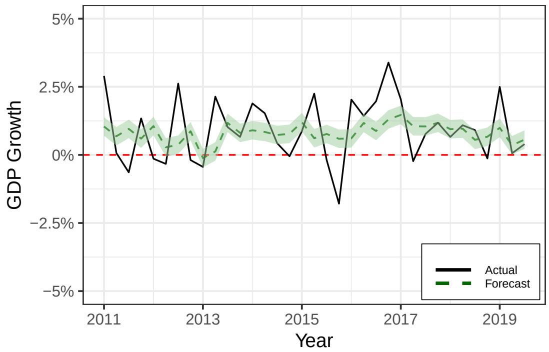
(a) BART



(b) GBM



(c) GLMNET



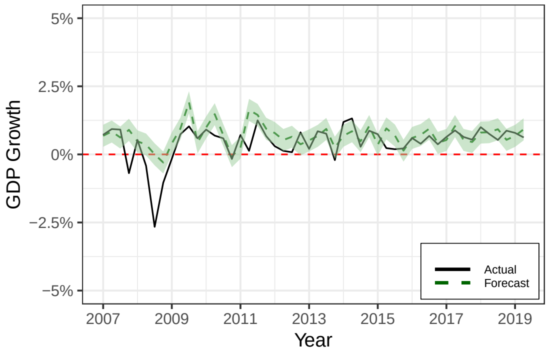
(d) XGBoost

Figure 5.8: ML forecasts - Great Britain

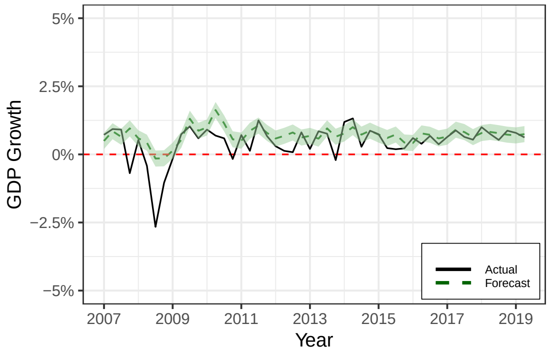
Note: The green shadow indicates the area one standard deviation from the forecast point value

5.2.5 USA

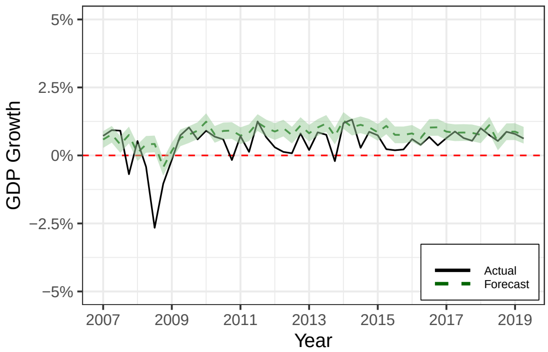
The time series of USA is particularly interesting, as the forecasts includes the Financial Crisis that occurred in 2008. As we can see, all of the TS regression models do to some extent forecast the sharp downturn in 2008. However, none of the models forecast the magnitude of the Financial Crisis. Of the TS regression model, the ARIMA model is the one with the best forecast performance. This indicates that the properties of the ARIMA model is relatively well suited for forecasting the GDP growth of USA. On the other hand, while the forecast accuracy is relatively adequate, all of the other ML algorithms yield lower RMSE values. Both the VAR models are yielding higher RMSE, despite the fact that the model includes key macroeconomic variables. This indicates that other variables might be more suited in building VAR models for forecasting the GDP growth of USA.



(a) AR

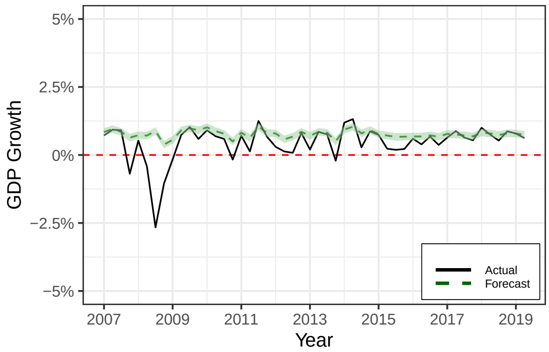


(b) ARIMA



(c) VAR_1*

* VAR_1 is a multivariate model containing GDP growth, growth rate of inflation and growth rate of unemployment rate



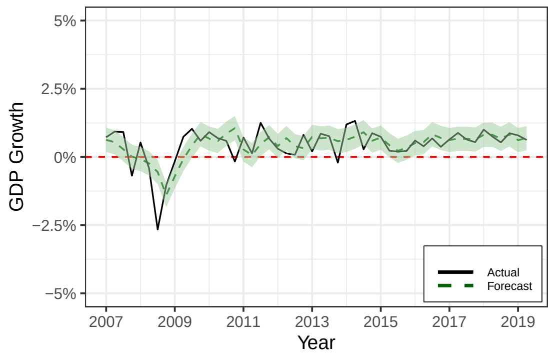
(d) VAR_2**

** VAR_2 is a multivariate model containing GDP growth, growth rate of inflation and interest rate

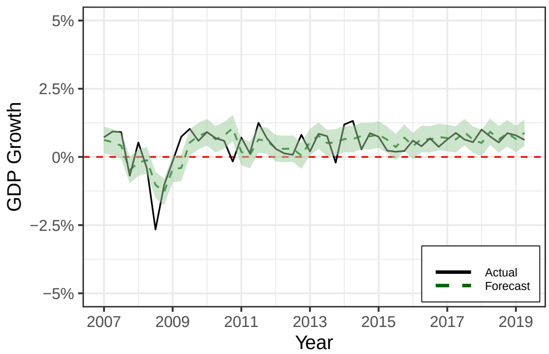
Figure 5.9: TS model forecasts - USA

Note: The green shadow indicates the area one standard deviation from the forecast point value

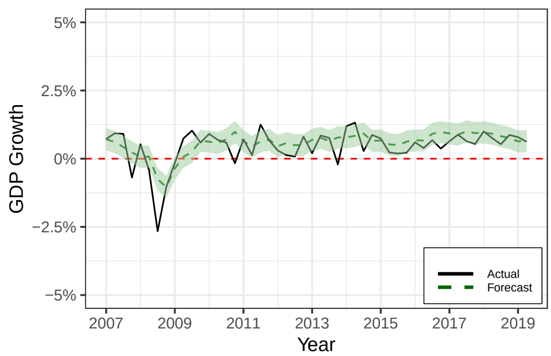
From Figure 5.10 we note that all of the ML algorithms successfully forecasts the financial crisis with a sharp downturn. However, as for the TS regression model, none of the algorithms are able to capture the magnitude of the downturn. This may suggest that the ML algorithms to an increased degree have the ability to work as an early warning system for financial crises than TS regression models. Compared to the forecast performance of all the models, including the TS regression models, the XGBoost algorithm yield best forecast performance. The tree variables that drives the forecasts of the XGBoost algorithm is a weighted average of 85 existing monthly indicators of national economic activity, percentage of survey respondents in the manufacturing sector who report improving business conditions and growth rate of food prices in the United States. Taking these variables into consideration, especially the first two variables may give a good indication of where the economy is heading. For instance, during the financial crisis the percentage of survey respondents in the manufacturing sector who report improving business conditions and growth rate of food prices in the United States. Taking these variables into consideration, especially the first two variables may give a good indication of where the economy is heading. For instance, during the financial crisis the percentage of survey respondents in the manufacturing sector who reported improving business conditions fell. The development of this variable, combined with the weighted average of 85 existing monthly indicators of national economic activity, made the XGBoost algorithm forecast the sharp downturn in GDP growth during the financial crisis. For the time series of USA, we see that for a particular long time series containing sufficent number of variables and observations the ML algorithms outperform the traditional TS regression models.



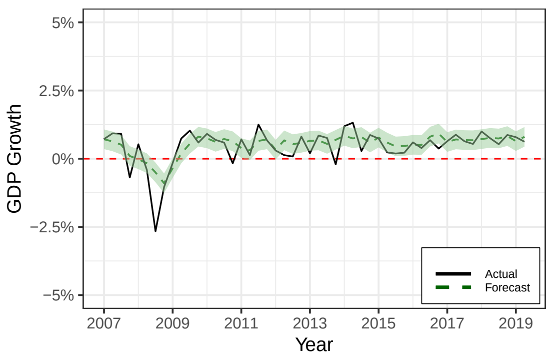
(a) BART



(b) GBM



(c) GLMNET



(d) XGBoost

Figure 5.10: ML forecasts - USA

Note: The green shadow indicates the area one standard deviation from the forecast point value

5.3 Variable assessment

In order to contrast ML algorithms with suggestions from macroeconomic theory, it is useful to do a variable assessment. During the last subchapter, we examined which variables that drives the forecasts of the ML algorithms for the selected economies. As mentioned in subsection 4.3.3, the literature suggests that unemployment rate and inflation rate are the best predictors for forecasting GDP growth. Moreover, the inclusion of interest rate rather than unemployment rate also showed to yield increased forecast performance compared to the ML algorithms. It is therefore interesting to investigate which variables that drives the algorithms for each country. We will do an assessment of the top three variables that drives the XGBoost algorithm for each country. The reason for only assessing the variable importance of the XGBoost is that this algorithm on average delivered the best forecast performance.

The Table 5.2 shows the top three variables that drives the forecasts of the XGBoost algorithm for the different economies. We notice that different versions of unemployment and inflation are important variables. For instance, for Canada the part time employment change seems be the most important variable for the XGBoost forecasts. However, it is hard to argue from a macroeconomic standpoint that this variable should be the most important for forecasting GDP growth, and this could be merely be a result of spurious correlation that the algorithm has detected from the Canadian time series. When it comes to the Euro Area we notice that the most important variable for forecasting GDP growth is long term unemployment rate. This variable is calculated as the fraction of eligible workers who have been unemployed and looking for work for a period of 12 months. Having this as the most important variable for forecasting GDP growth seems more intuitive. The link between GDP and unemployment rate was first proposed by Okun (1965), whose research stated that every 1% increase in unemployment rate would lead to lower a country's potential GNP with 2%. While the unemployment rate of part time employees also would affect the potential GNP, the workforce of part time employees is naturally smaller than for the work force as a whole. Furthermore, when examining the results of the forecasts of the Swedish GDP growth, the most important is the long term unemployment rate also in this case. Seeing this in relation to the forecast accuracy of XGBoost model, we notice the XGBoost has highest forecast performance of all the ML algorithms for both Euro Area and Sweden.

Country	Variable
Australia	1. Consumer Confidence 2. Manufacturing Production 3. Core Inflation Rate
Canada	1. Part Time Employment Change 2. Producer Prices Change 3. Retail Sales MoM
Germany	1. Factory Orders 2. Inflation Rate MoM 3. Car Registrations
Spain	1. Retail Sales YoY 2. New Orders 3. Loans to Private Sector
Euro Area	1. Long Term Unemployment Rate 2. Zew Economic Sentiment Index 3. Inflation Rate
France	1. Hourly Wages in Industry Index 2. Zew Economic Sentiment Index 3. Government Bond 10Y
Japan	1. Consumer Spending 2. Consumer Confidence 3. Car Registrations
Sweden	1. Long Term Unemployment Rate 2. Retail Sales YoY 3. Manufacturing PMI
United Kingdom	1. Industrial Production 2. Bankruptcies 3. Balance of Trade
USA	1. Chicago Fed National Activity Index 2. ISM Purchasing Managers Index (PMI) 3. Food Inflation

Note: Description of each variable including units and sources may be found here (external link)

Table 5.2: Variable importance drawn from XGBoost algorithm

Moreover, in the results of the XGBoost models for Australia, Canada, Germany, Euro Area and USA variables consisting of different versions of inflation rate are important. For instance, for Australia the inflation rate is the rate of change of the core Consumer Price Index, while for Canada the algorithm emphasises monthly percentage change in the Producer Prices Index. For the United States the XGBoost emphasizes the growth rate of food prices. It is also interesting that the number of car registrations, which measures the number of new car registrations per month, is of such importance for the forecasts of GDP growth in Germany and Japan. This particular finding is interesting because of two reasons. Firstly, the automotive industry are both prominent and large industry for the respective economies. In 2018, Japan and Germany was respectively the third and fifth largest producers of motor vehicles in the world (OICA, 2018). Secondly, assuming that owning a car is a luxury good, the number of new car registrations might reveal some information about the consumer confidence in the economy. Drawing on classical microeconomic theory, consumers tend to buy more luxury goods when their income increases while consumers tend to buy more inferior goods when income levels drops. While we do find some evidence of macroeconomic theory in the variables that drive the XGBoost algorithm, we also find some interesting uncommon predictors.

6 Discussion

The future is uncertain and forecasting the future is thus inherently difficult. The results in this thesis indicate that traditional TS regression models are better suited than ML algorithms to forecast GDP growth. The superior forecasting abilities of the simpler TS regression models brings to mind the principle of Occam's razor. The principle originally stems from philosophy and states that the more assumptions you have to make for an occurrence, the more unlikely the explanation. Green and Armstrong (2015) researched the predictive validity of the principle, and found that none of 97 comparisons in 32 papers provided a balance of evidence that complexity increases forecast accuracy. Therefore, it is worth questioning whether the devotion of considerable resources in order to attempt to make sophisticated forecasting models is at all useful. It seems rather that forecasting is most useful, and least prone to bias, when limited to simple, applicable models. When using simple forecasting models, there is also greater probability that the utility of the forecasting exceeds the cost of forecasting.

Although the results indicate the superiority of traditional TS regression models, this does not necessarily mean that ML algorithms will never produce better forecasts than traditional TS regression models. ML algorithms normally require a lot of data to learn complex patterns and relationships between variables. For a particular long time series with a rich number of variables (e.g. USA), we see that the ML algorithms outperform the traditional TS regression models. In addition, while the algorithms underestimates the magnitude of the Financial Crisis in 2008, they all correctly forecasts the sharp downturn in the economy. This indicates when provided with sufficient number of observations and variables, the algorithms are able to pick up the variables that drives GDP growth. The implication is that employing ML algorithms on forecasting macroeconomic variables may be more relevant and produce more adequate results, when we have access to the sufficient amount of data.

Another important aspect is whether the included variables for each time series correctly contains all the factors that might affect GDP growth. For instance, we might have gathered those variables that are important for GDP growth in USA, but excluded variables that might be important for UK or those variables might not exist at the moment. As a result, the ML algorithms might be training on variables that may not be useful in forecasting GDP growth. Furthermore, we noticed for instance that the GBM model predicted a recession to occur when the true GDP growth was at peak level for the Canadian time series. The underlying variables that drove the GBM model indicated that the GDP growth should have been declining, but the true GDP growth was increasing. This indicates that the variables included for the Canadian time series, may not explain all of the variation in what affects the Canadian GDP growth. Therefore we cannot say whether forecast accuracy performance of the ML algorithms are due to the algorithm itself or the lack of access to those variables that explains the true variation of GDP growth.

ML algorithms is severely dependent on the amount and quality of data. With the time, we would most likely get access to more relevant macroeconomic variables that will lead to better forecasts using ML algorithms. However, for the time being our results indicate that traditional TS regression models in general produce more accurate forecasts than ML algorithms and we have experienced the relevance of Occam's razor.

7 Conclusion

Applying ML algorithms on forecasting macroeconomic variables is a new and growing field, and has shown to be inherently suitable in the finance, healthcare and retail industry. We therefore employ state-of-the-art ML algorithms on forecasting GDP growth and compare the results to traditional TS regression models. To increase the validity and generalization, we evaluate 10 different economies; *Australia, Canada, Euro Area, Germany, Spain, France, Japan, Sweden, Great Britain* and *USA*. Furthermore, we also do an assessment of which variables that drives the XGBoost algorithm, as it was the best performing ML model.

Our results show that the traditional TS regression models in general outperfom ML algorithms in terms of forecasting performance. The VAR_2, which is a multivariate model built on inflation rate and interest rate, on average yielded lowest RMSE and MDA. In 6/10 cases the VAR_2 model was able to outperform the majority of the included models in terms of anticipating the directional movement of the GDP growth, indicating the model's and the included variables' suitability of forecasting GDP growth. Furthermore, we see that the XGBoost algorithm have in most cases identified reasonable macroeconomic variables, but do in general propose different variables than what macroeconomic theory proposes are important predictors for GDP growth. The XGBoost algorithm might have produced higher forecasting accuracy, if it had put more emphasize interest rate and unemployment rate.

Even though the TS regression models in the clear majority of instances outperformed the forecasting performance of the ML algorithms, we suggest that at the moment there do not exist sufficient amount of macroeconomic variables and observations that the algorithms can learn from. For the largest time series of USA, we noticed that the ML algorithms delivered better forecasting performance. This indicates that we may employ ML algorithms, when sufficient amount of data becomes available. However, for the time being; less is more in forecasting GDP growth.

References

Andersson, J. (2007). Forecasting swedish gdp growth.

Bernshtein, N. A. (1967). *The co-ordination and regulation of movements.* Oxford: Pergamon Press.

Canova, F. (1995). Vector autoregressive models: specification, estimation, inference and forecasting. *Handbook of applied econometrics*, 1:73–138.

Chakraborty, C. and Joseph, A. (2017). Machine learning at central banks.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Christiano, L. J., Eichenbaum, M., and Evans, C. L. (1999). Monetary policy shocks: What have we learned and to what end? *Handbook of macroeconomics*, 1:65–148.

Din, M. (2015). Arima by box jenkins methodology for estimation and forecasting models in higher education.

Economist, T. (2015). How machine learning works.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Green, K. C. and Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research*, 68(8):1678–1685.

Greenwell, B. (2019). Hands-on machine learning with r.

Hastie, T. and Qian, J. (2014). Glmnet vignette. *Retrieve from http://www. web. stanford. edu/˜ hastie/Papers/Glmnet_Vignette. pdf. Accessed September*, 20:2016.

Hilde, C. B. and Thorsrud, L. A. (2014). *Applied Time Series For Macroeconomics.* Gyldendal Norsk Forslag AS, 1 edition.

Hyndman, R. J., . K. (2008). Automatic time series forecasting: The forecast package for r. journal of statistical software. *Journal of Statistical Software*, 27(1):1–22.

Jones, C. I. (2014). *Macroeconomics.* New York : W.W. NORTON Company, 3 edition.

Jung, J.-K., Patnam, M., and Ter-Martirosyan, A. (2018). *An Algorithmic Crystal Ball: Forecasts-based on Machine Learning.* International Monetary Fund.

Kapelner, A. and Bleich, J. (2013). bartmachine: Machine learning with bayesian additive regression trees. *arXiv preprint arXiv:1312.2171.*

Marcellino, M., Stock, J. H., and Watson, M. W. (2006). A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of econometrics*, 135(1-2):499–526.

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal.

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21.

Nations, U. (2010). *System of National Accounts 2008*.

OICA (2018). International organization of motor vehicle manufacturer - 2018 production statistics.

Okun, A. M. (1965). The gap between actual and potential output. *The Battle against Unemployment*.

Richardson, A., Mulder, T., et al. (2018). Nowcasting new zealand gdp using machine learning algorithms.

Saman, C. (2011). Scenarios of the Romanian GDP Evolution With Neural Models. *Journal for Economic Forecasting*, 0(4):129–140.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (1970). Dropout: A simple way to prevent neural networks from overfitting.

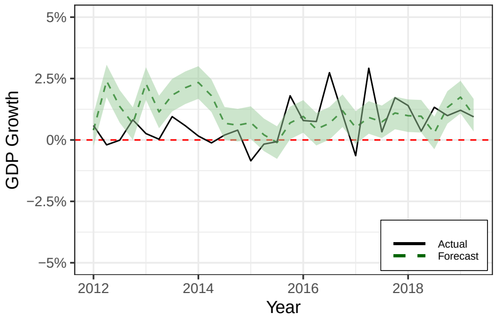
Stoyanov, V., Ropson, A., and Eisner, J. (2011). Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 725–733.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International journal of forecasting*, 16(4):437–450.

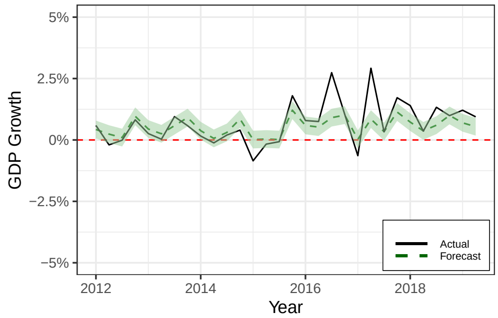
Tkacz, G. (2001). Neural network forecasting of canadian gdp growth. *International Journal of Forecasting*, 17(1):57–69.

Appendix

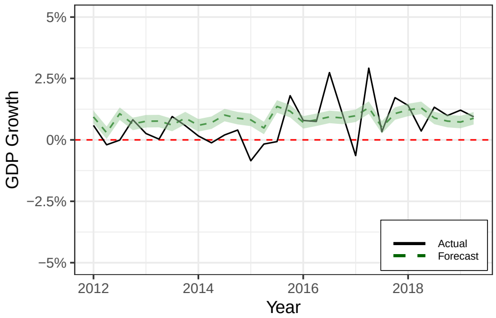
A1 Australia



(a) AR

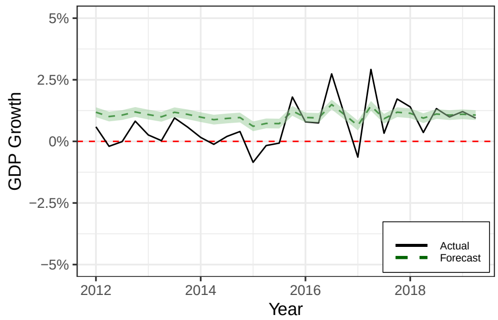


(b) ARIMA



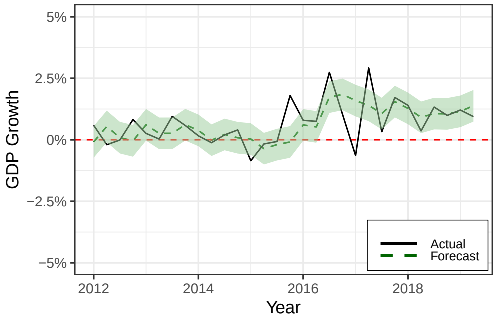
(c) VAR_1*

* VAR_1 is a multivariate model containing GDP growth and growth rate of inflation and unemployment

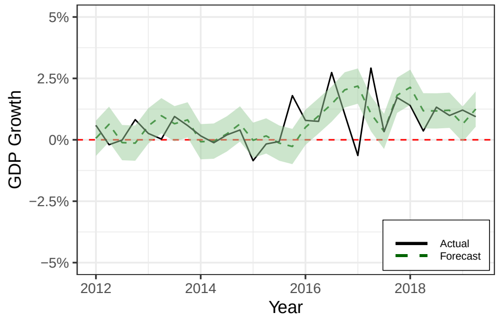


(d) VAR_2**

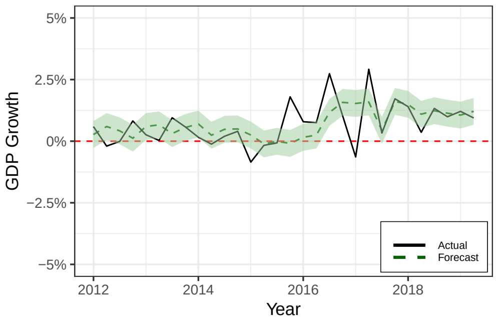
** VAR_2 is a multivariate model containing GDP growth, growth rate of inflation and interest rate



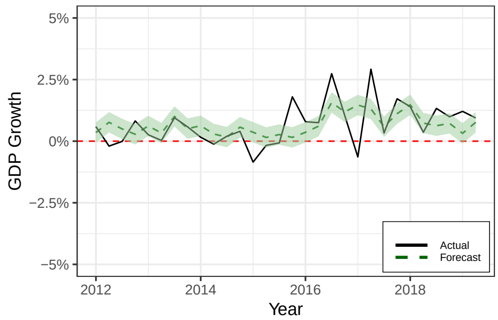
(e) BART



(f) GBM



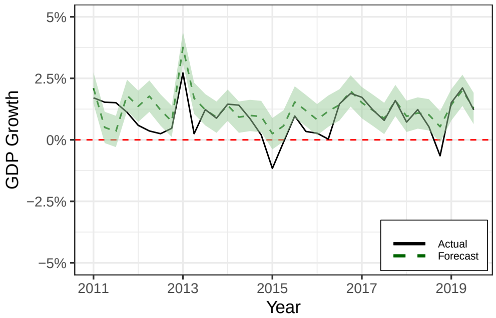
(g) GLMNET



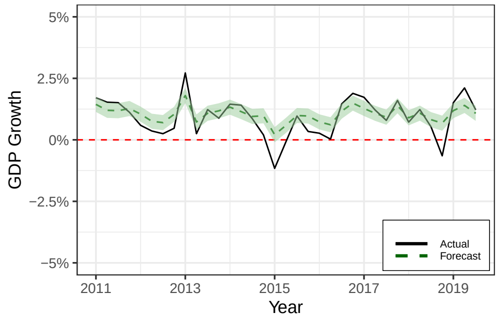
(h) XGBoost

Note: The green shadow indicates the area one standard deviation from the forecast point value

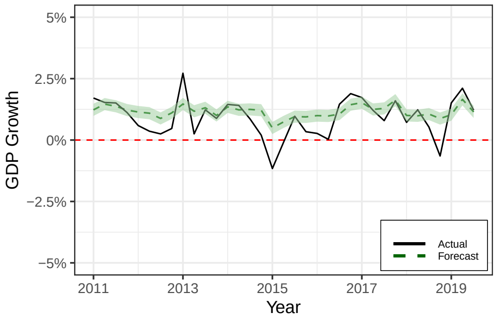
A2 Canada



(a) AR

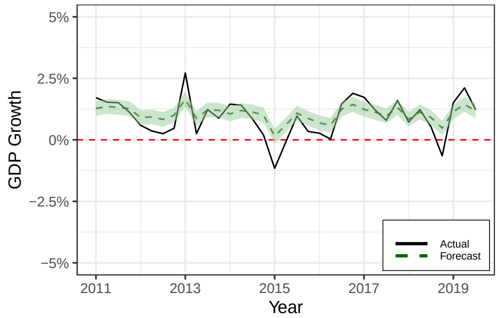


(b) ARIMA



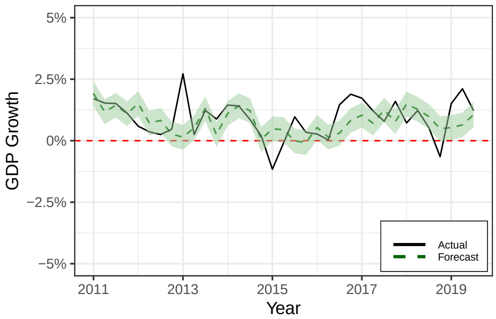
(c) VAR_1*

* VAR_1 is a multivariate model containing GDP growth and growth rate of inflation and unemployment

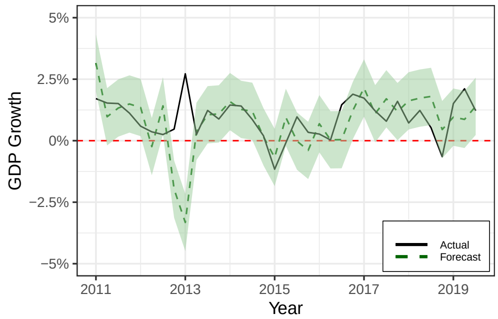


(d) VAR_2**

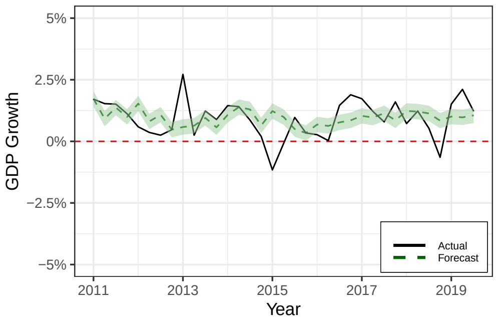
** VAR_2 is a multivariate model containing GDP growth, growth rate of inflation and interest rate



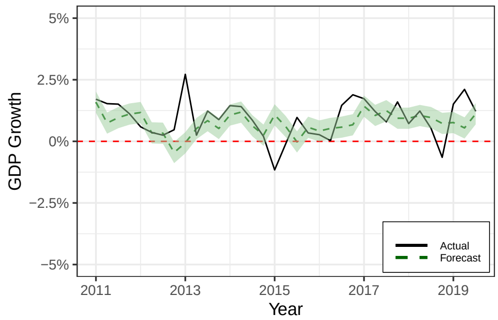
(e) BART



(f) GBM



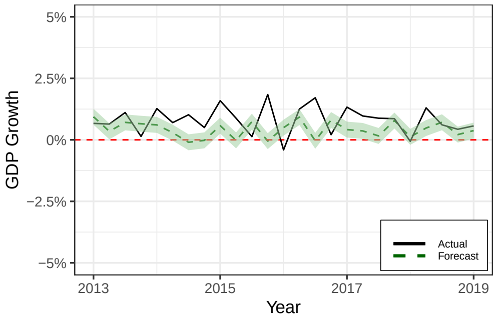
(g) GLMNET



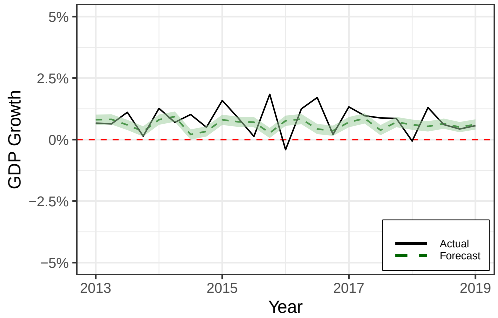
(h) XGBoost

Note: The green shadow indicates the area one standard deviation from the forecast point value

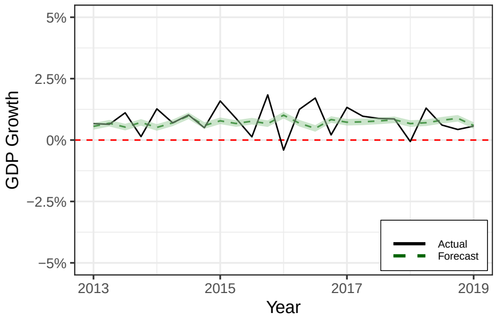
A3 Germany



(a) AR

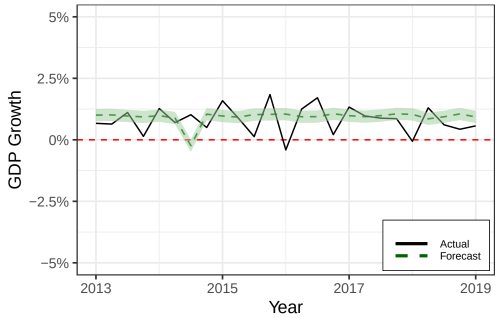


(b) ARIMA



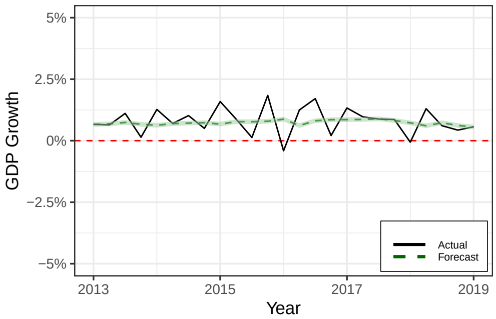
(c) VAR_1*

* VAR_1 is a multivariate model containing GDP growth and growth rate of inflation and unemployment

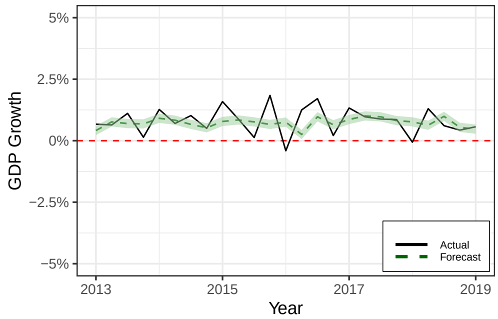


(d) VAR_2**

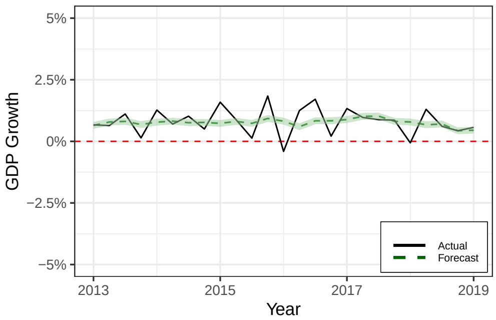
** VAR_2 is a multivariate model containing GDP growth, growth rate of inflation and interest rate



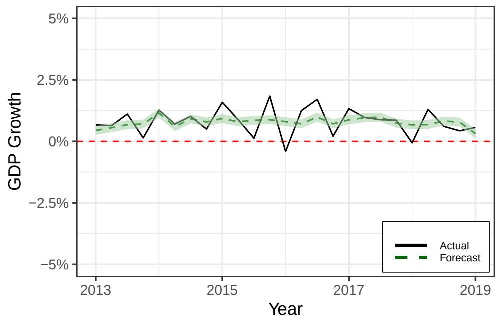
(e) BART



(f) GBM



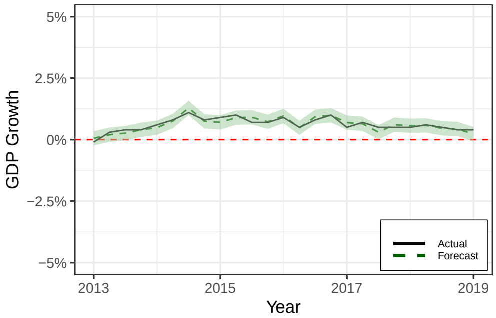
(g) GLMNET



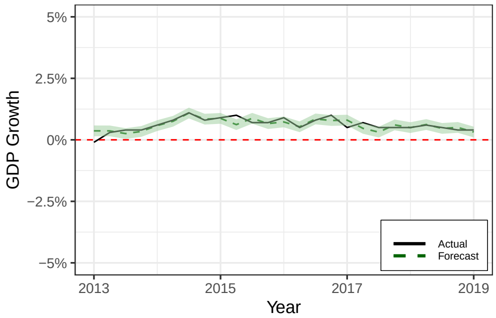
(h) XGBoost

Note: The green shadow indicates the area one standard deviation from the forecast point value

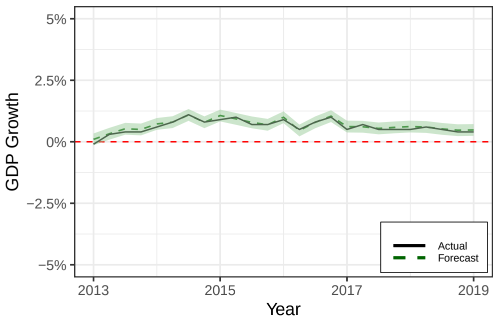
A4 Spain



(a) AR

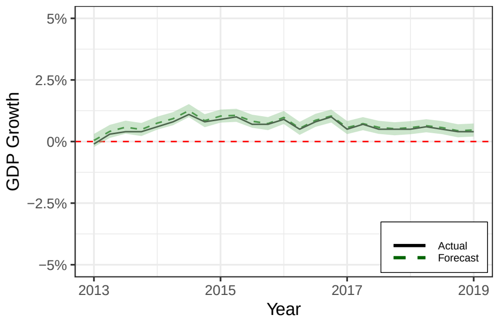


(b) ARIMA



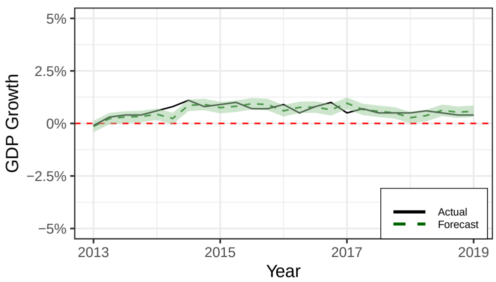
(c) VAR_1*

* VAR_1 is a multivariate model containing GDP growth and growth rate of inflation and unemployment

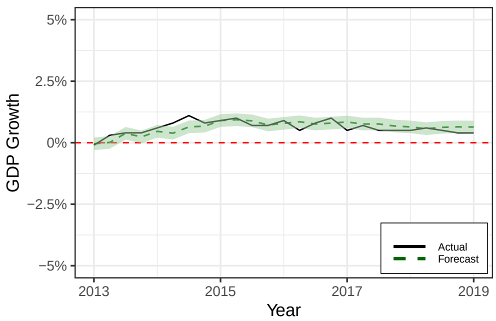


(d) VAR_2**

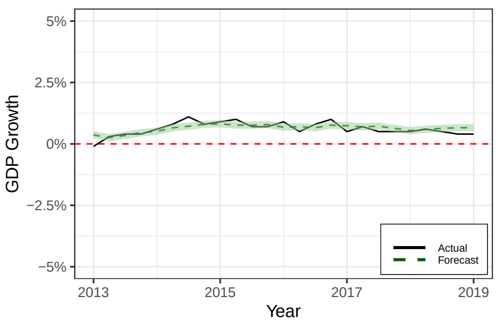
** VAR_2 is a multivariate model containing GDP growth, growth rate of inflation and interest rate



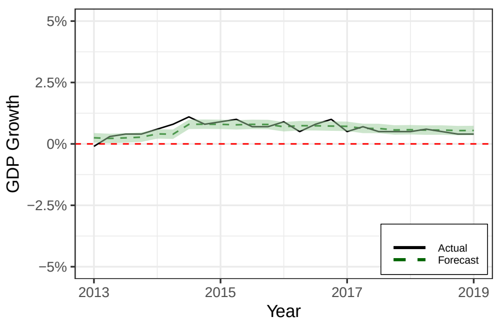
(e) BART



(f) GBM



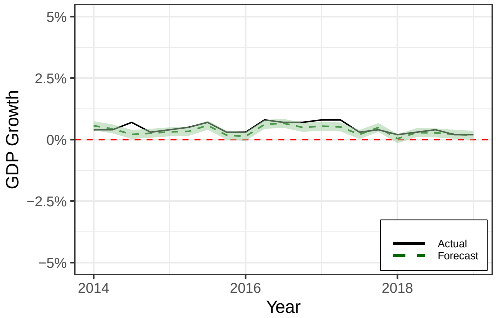
(g) GLMNET



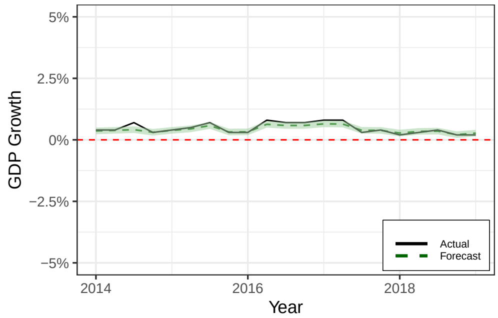
(h) XGBoost

Note: The green shadow indicates the area one standard deviation from the forecast point value

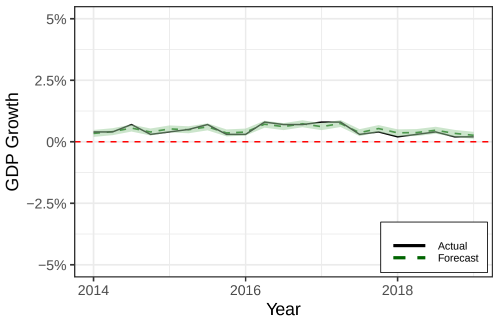
A5 Euro Area



(a) AR

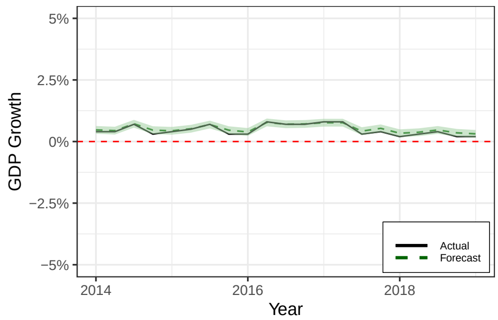


(b) ARIMA



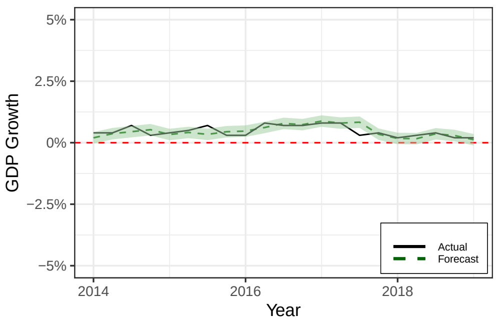
(c) VAR_1*

* VAR_1 is a multivariate model containing GDP growth and growth rate of inflation and unemployment

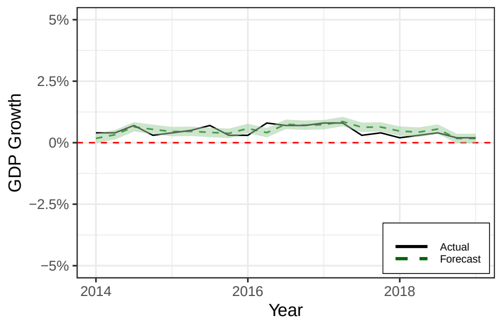


(d) VAR_2**

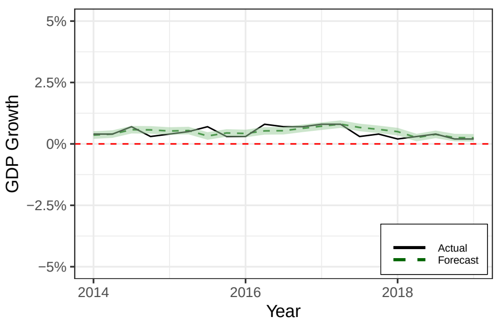
** VAR_2 is a multivariate model containing GDP growth, growth rate of inflation and interest rate



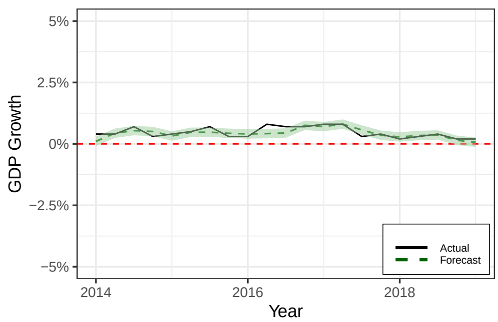
(e) BART



(f) GBM



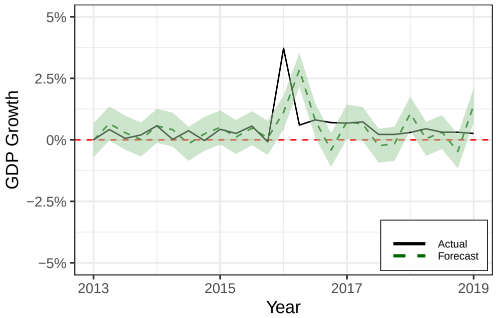
(g) GLMNET



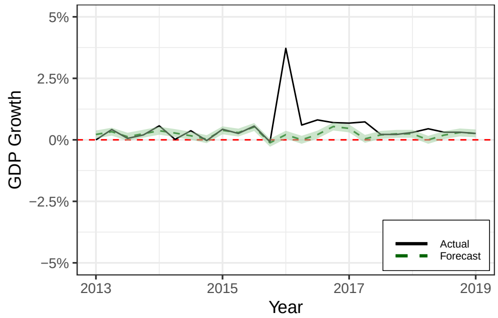
(h) XGBoost

Note: The green shadow indicates the area one standard deviation from the forecast point value

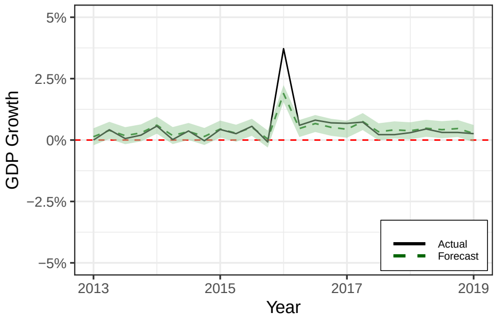
A6 France



(a) AR

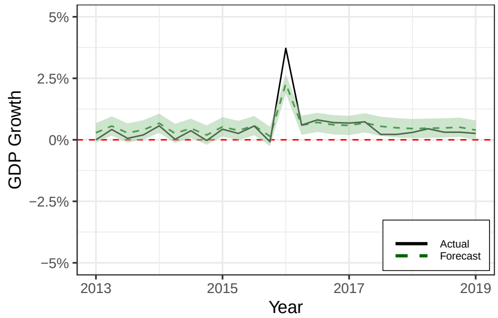


(b) ARIMA



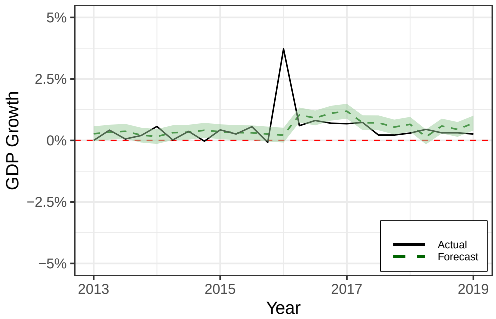
(c) VAR_1*

* VAR_1 is a multivariate model containing GDP growth and growth rate of inflation and unemployment

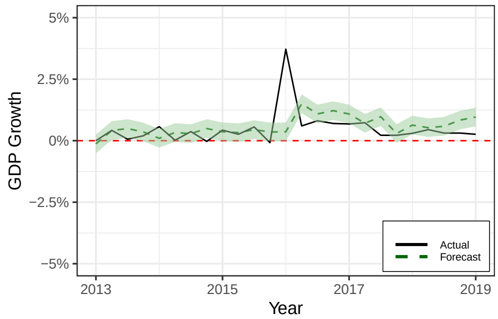


(d) VAR_2**

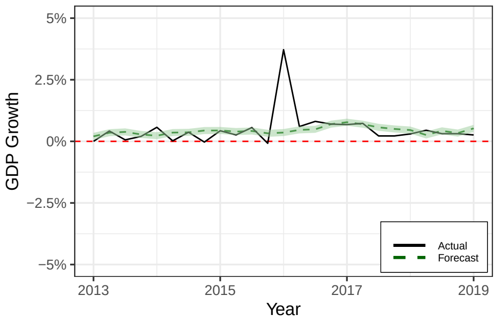
** VAR_2 is a multivariate model containing GDP growth, growth rate of inflation and interest rate



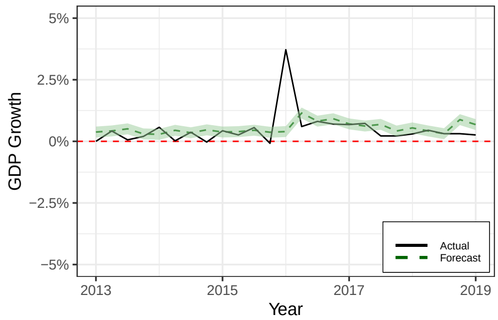
(e) BART



(f) GBM



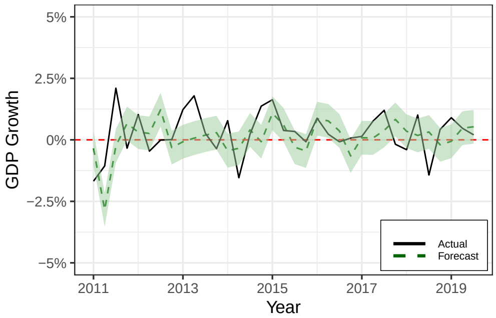
(g) GLMNET



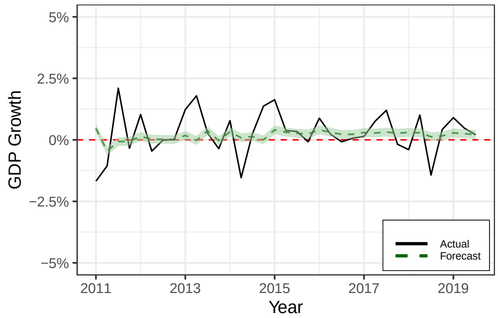
(h) XGBoost

Note: The green shadow indicates the area one standard deviation from the forecast point value

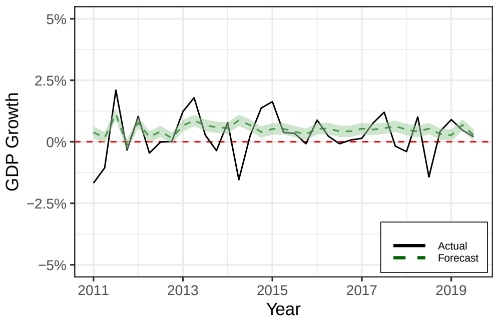
A7 Japan



(a) AR

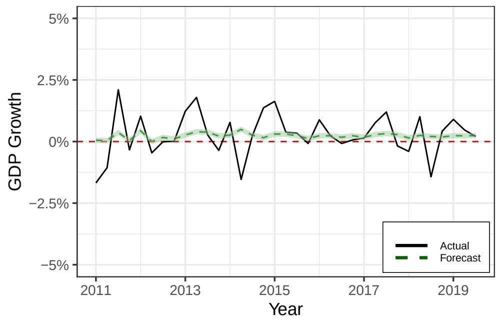


(b) ARIMA



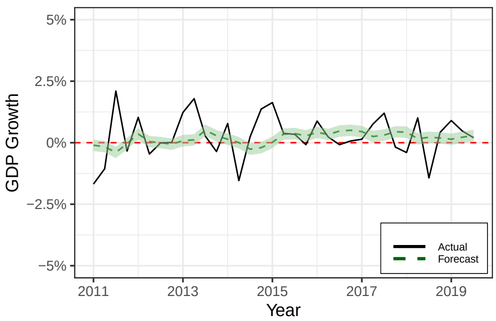
(c) VAR_1*

* VAR_1 is a multivariate model containing GDP growth and growth rate of inflation and unemployment

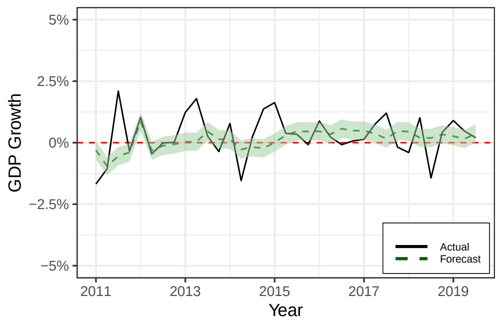


(d) VAR_2**

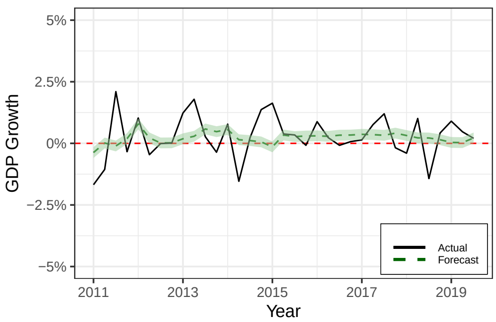
** VAR_2 is a multivariate model containing GDP growth, growth rate of inflation and interest rate



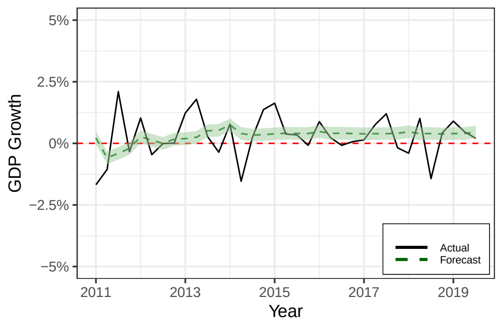
(e) BART



(f) GBM



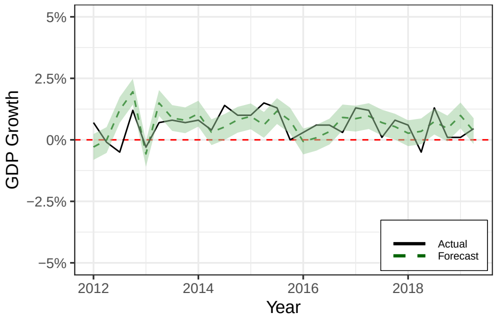
(g) GLMNET



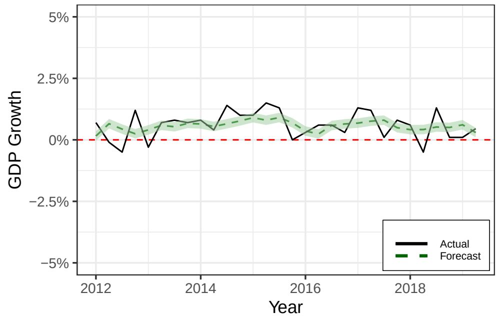
(h) XGBoost

Note: The green shadow indicates the area one standard deviation from the forecast point value

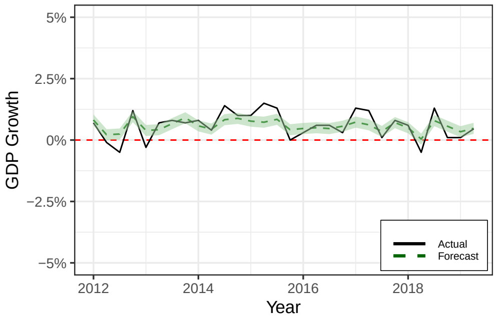
A8 Sweden



(a) AR

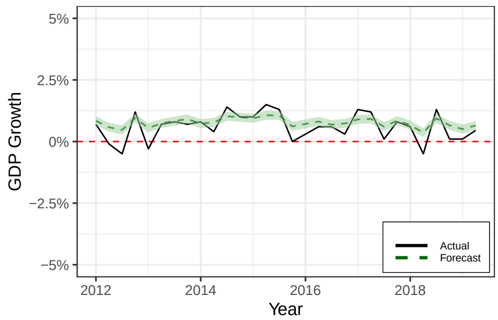


(b) ARIMA



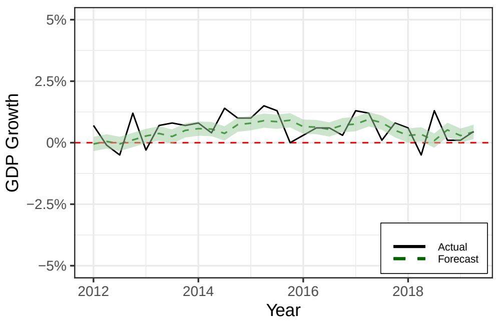
(c) VAR_1*

* VAR_1 is a multivariate model containing GDP growth and growth rate of inflation and unemployment

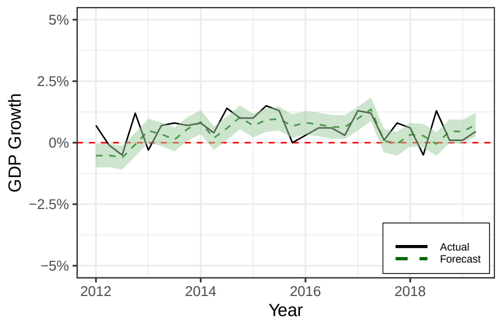


(d) VAR_2**

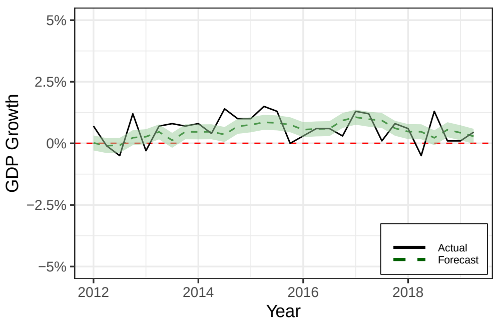
** VAR_2 is a multivariate model containing GDP growth, growth rate of inflation and interest rate



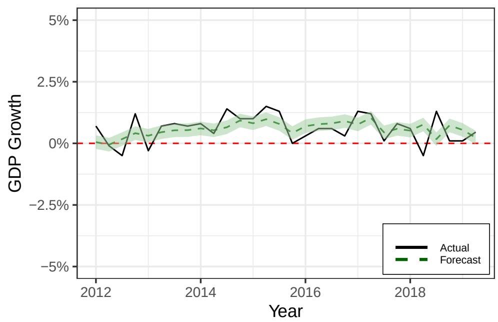
(e) BART



(f) GBM



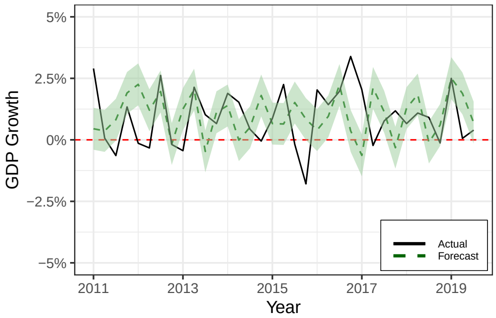
(g) GLMNET



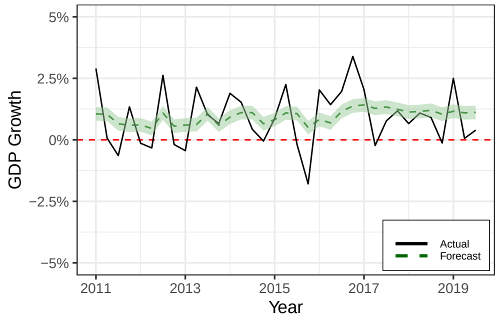
(h) XGBoost

Note: The green shadow indicates the area one standard deviation from the forecast point value

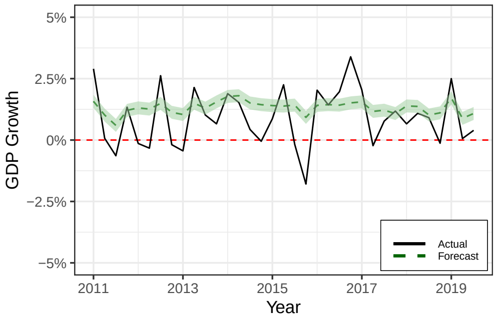
A9 Great Britain



(a) AR

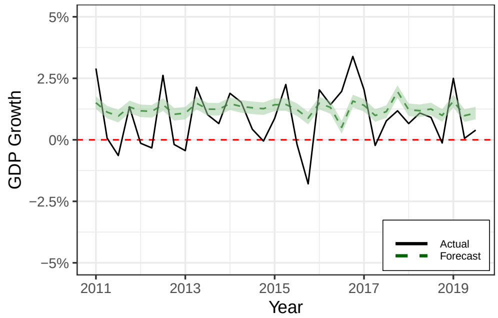


(b) ARIMA



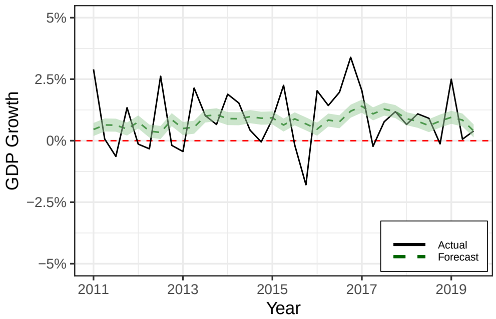
(c) VAR_1*

* VAR_1 is a multivariate model containing GDP growth and growth rate of inflation and unemployment

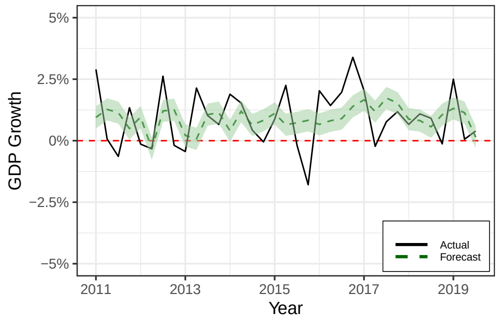


(d) VAR_2**

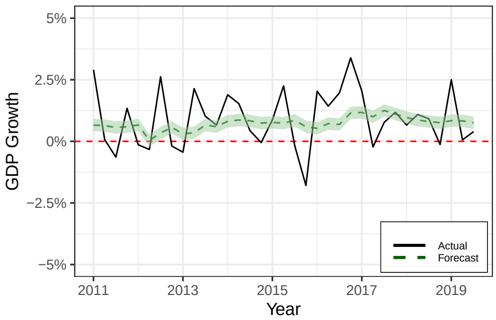
** VAR_2 is a multivariate model containing GDP growth, growth rate of inflation and interest rate



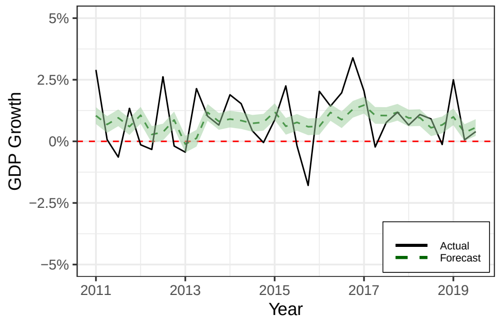
(e) BART



(f) GBM



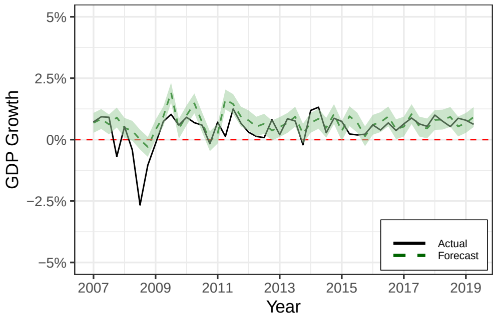
(g) GLMNET



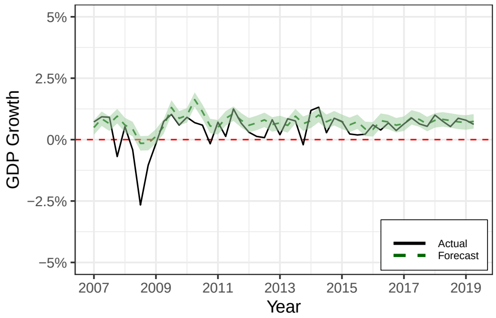
(h) XGBoost

Note: The green shadow indicates the area one standard deviation from the forecast point value

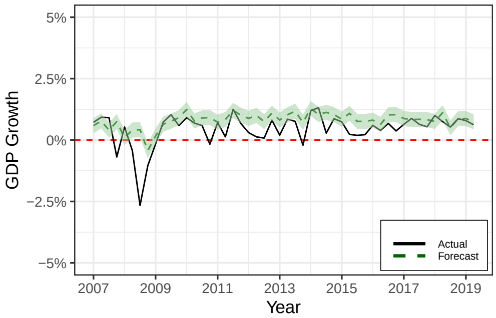
A10 USA



(a) AR

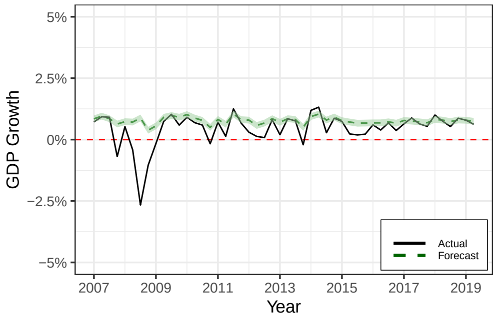


(b) ARIMA



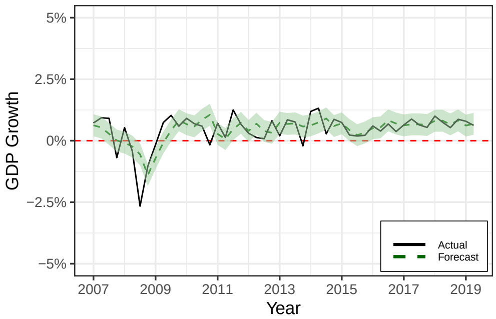
(c) VAR_1*

* VAR_1 is a multivariate model containing GDP growth and growth rate of inflation and unemployment

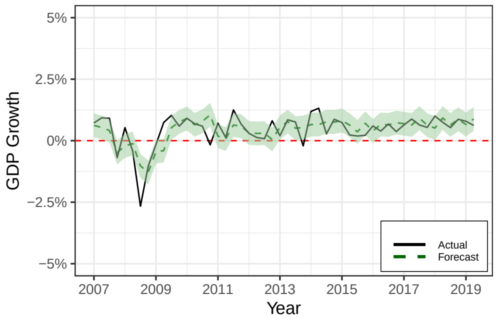


(d) VAR_2**

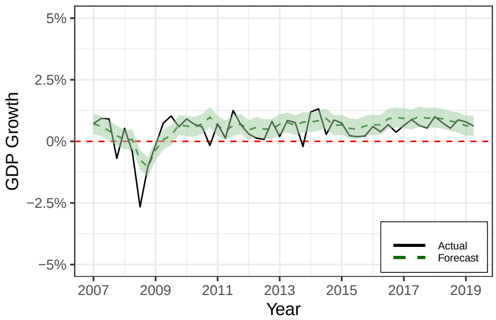
** VAR_2 is a multivariate model containing GDP growth, growth rate of inflation and interest rate



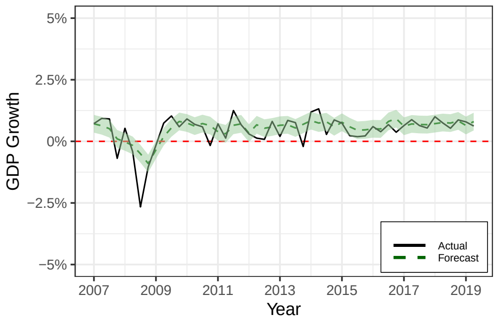
(e) BART



(f) GBM



(g) GLMNET



(h) XGBoost

Note: The green shadow indicates the area one standard deviation from the forecast point value