Nowcasting of the economic dynamic in the Dominican Republic under the Covid-19 crisis: An approach based in Big Data and machine learning techniques[1]

Juan Salvador Quiñonez Wu[2] Lisette Josefina Santana Jimenez[3]

Abstract

The availability of a huge amount of information, both quantitative and qualitative, structured and unstructured (big data), jointly with the emergence of machine learning techniques, appropriate for its treatment, has led to a revolution in various areas of knowledge, extending the data pool and analytical framework for decision-making processes. In this sense, it has become possible to generate answers to various research questions that could not be addressed under a traditional empirical approach. Macroeconomic modeling has been considerably favored by the big data-machine learning binomial, mainly under the prevailing situation of the Covid-19 crisis, where the speed and greater scope of decision-making processes have appealed to the use of granular information, available in real time. The objective of this document is to present a set of models, based on machine learning techniques, used to carry out nowcasting and generate leading indicators of the economic activity in the Dominican Republic. Through the system used (composed of a bayesian structural model of time series, a multi-layer perceptron and lasso and ridge regressions) the forecast errors for the monthly index of economic activity (IMAE) are minimized (measured in terms of the root mean squared error, RMSE) and has better performance than the benchmark, which is an autoregressive integrated moving average (ARIMA) model. The potential of these techniques is emphasized in terms of macroeconomic modeling, to assess the balance of risks of certain variables and for the decision-making processes, even in non-linear or chaotic scenarios.

Key words: Big Data, machine learning, economic dynamic, monthly index of economic activity.

JEL classification: E21, E22, E23, E27, C59.

[1] The opinions expressed in this document are responsibility of the authors and do not necessarily represent the posture of the Central Bank of the Dominican Republic.

[2] Central Bank of the Dominican Republic. University of Utah, United States.

[3] Central Bank of the Dominican Republic. University of Warwick, United Kingdom.

I. Introduction

The objective of the nowcasting exercises is to make very short-term projections on the behavior of certain variables, using a set of data available with higher frequency than the dependent variable, making a timelier analysis of its expected trajectory and the balance of risks associated with it. It is estimated that, daily, the humanity produces 2.5 exabytes of information, of which 90% has been generated in the last two years, and it is expected that, by 2025, these values will continue to increase exponentially, reaching 463 exabytes, globally. Real-time access to high-dimensional data sets (big data) has not only played a leading role in enriching the empirical literature concerning nowcasting, but also the availability of this enormous amount of information has facilitated decision-making processes, including the management of monetary policy.

The nature of these data sets, in which the number of features is greater than the amount of observations, does not admit a treatment under traditional econometric approaches, given the restrictions of these techniques to deal with high-dimensional information, but rather under heuristic approaches. which do not incur in problems related to the curse of dimensionality[4] and that provide greater flexibility, making optimal use of the available information (Makridakis *et al,* 2018). In this sense, the economic literature has appealed to machine learning techniques, which have become more relevant, considering the benefits they offer to manage these data and minimize forecast errors, even under non-linear and chaotic scenarios.

Under the Covid-19 crisis, timely access to different sources of information (including non-conventional sources) became increasingly imminent, given the speed required by different decision-making processes at a multi-dimensional level, including the management of monetary policy. In this sense, the compilation of high-frequency granular information (quantitative and qualitative; structured/unstructured), as well as the emergence of appropriate tools for the treatment of this data, have played a preponderant role in generating projections and evaluating the balance of risks associated with the main macroeconomic variables, as well as to build different metrics linked to uncertainty.

In this way, both the access to a greater stock of information and the timely projection of the potential trajectory of the main macroeconomic and financial variables constitute, from the perspective of

[4] The curse of dimensionality refers to the sub-optimal performance of certain models and algorithms when dealing with high dimensional data sets.

central banks, an added value to manage monetary policy more efficiently and to optimize its impact at the level of the different productive sectors.

The objective of this document is to present a set of models, based on machine learning techniques, used to carry out nowcasting and generate leading indicators of economic activity in the Dominican Republic. The projections are generated for the monthly index of economic activity (IMAE), using a bayesian structural time series model (BSTS, Scott & Varian, 2017), multi-layer perceptron model (MLP) and lasso and ridge regressions.

The results of the out-of-sample projections, at a time horizon $h = 1$, are measured in terms of the root mean square error (RMSE) is relatively minimum, compared to the model chosen as benchmark (*i.e.* autoregressive integrated moving average, ARIMA). The potential of the binomial big data and machine learning is emphasized and it is established how this type of techniques are emerging as tools with great performance in terms of macroeconomic modeling, minimizing forecast errors, even in non-linear or chaotic scenarios, and expanding the analytical framework for decision-making processes, as well as for the evaluation of the balance of risks of different variables.

The structure of this document is as follows: in the subsequent section, a review of the empirical literature concerning the use of machine learning models to project macroeconomic and financial variables is carried out; section III presents the information used in the models considered for each variable as well as the empirical approach used to obtain the results of this document; the results of the forecasting exercise and the comparison with the benchmark model are shown in section IV; finally the conclusions reached in this paper are discussed in section V.

II. Literature Review

The economic literature concerning the use of machine learning (ML) models and algorithms is relatively recent but it has been increasing exponentially, considering the advantages of these techniques for the treatment of high dimensional data, minimizing the forecasting errors, even in non-linear and chaotic scenarios. It is also important to emphasized that these techniques allow to extract information both from multivariate quantitative datasets and qualitative sources of information, extending the access to alternative platforms of data and enriching the analytical framework for the decisions of the policy makers.

Dauphin et al (2022) describe the most recent work aimed at strengthening the nowcasting exercises of the European affairs department of the International Monetary Fund (IMF). A set of variables from both conventional sources and non-conventional platforms (*e.g.* Google searches, air quality) are used. A standard model of dynamic factors and various ML algorithms are used to generate the nowcast of the growth of the Gross Domestic Product (GDP) for a heterogeneous group of European economies, both in ordinary periods and in crisis scenarios. It is verified that most of the techniques inherent to ML show a considerably higher performance than that of an AR (1) model, used as a point of comparison (benchmark model); it is emphasized that ML algorithms maintain this predictive power even under periods of high uncertainty, mainly to identify turning points in the series.

Richardson et al (2019) evaluate the performance of a set of ML algorithms in real time, in order to obtain more accurate GDP forecasts for the case of New Zealand; the authors estimate various ML models for the period 2009-2019, using large data sets (including domestic and foreign variables) and, subsequently, make comparisons with a naive auto-regressive model and with a dynamic factor model. The results found suggest that the precision of the projections of the Reserve Bank of New Zealand, in different periods, could have been optimized with ML algorithms.

Santana (2018) studies the potential of the big data-ML binomial to generate nowcast in the case of the tourism sector in the Dominican Republic, using a combination of data from official sources and searches from the Google Trends platform. The treatment of this information is carried out using a bayesian structural time series model (BSTS, Scott & Varian, 2016) with spike & slab priors; through this model, a reduction of the dimensionality of the data set used is carried out. The results show superior performance to a state-space model, without a regression component, as well as to a vector auto-regressive (VAR), used for comparison purposes (benchmark models). In addition to minimizing the forecast error, the model used, together with the selected variables, allow the identification of turning points in the series considered.

Other works in this line of research focus on nowcasting to identify turning points in the economic cycle. Azqueta-Gavaldon (2020) compute network topology to create an extensive data set, which they evaluate using ML techniques. Once cyclical patterns are identified in the considered network, various classifiers (naive-bayes, support vector machine and logistic regression) are used to predict different economic regimes in real time (e.g. periods of expansion and recessions). Likewise, Garbellano (2016) focuses on performing nowcasts for the economic cycle, emphasizing the superior performance of the k-nearest neighbor (KNN) algorithm for the variable selection process and, at the

same time, pointing out the potential of the combination of different approaches inherent to ML to obtain a final forecast.

On the other hand, some authors use a dynamic factor model as an intermediate step to reduce the dimensionality of the data sets considered and generate projections from ML algorithms. Soybilgen and Yazgan (2021) carry out short-term projections for the annual growth rate of GDP in the United States, using ensemble regression trees; the results suggest that the models considered have a better performance in the period after the financial crisis of 2008. A comparison of the results with the GDPNow projections of the Atlanta Federal Reserve is carried out, identifying a greater precision of the ML models in the first half of the quarter; however, GDPNow is more accurate towards the end of the considered quarter.

In general terms, it is evident that the empirical literature inherent to the use of tools inherent to ML demonstrates a considerable evolution for multivariate analysis and continues to show a favorable performance, mainly under the prevailing situation (Covid-19 crisis) and given the speed of decision-making processes, considerably favored by the concatenation of large data sets with ML techniques.

III. Data & Methodology

3.1. Data

In the case of the BSTS model, the lasso and ridge regressions, the information used to generate the projections of the monthly index of economic activity (IMAE), was compiled from different sources of information: (i) macroeconomic variables from the Central Bank of the Dominican Republic (CBDR); (ii) commodities prices (Bloomberg), gasoline prices (Minister of Industry and Commerce and MSMEs); (iii) metadata from the Google Trends platform. The time period considered is January 2017-March 2022, on a weekly basis; a dataset of 276 variables (Annex 1A) is used. The time series proceeding from the Google Trends platform, which provides a normalized index of queries, depending on the relative inter-temporal popularity of each term; the maximum value is 100 and the minimum value is zero.

The identification of the keywords for the searches was carried out in two different ways: first, terms related to the acquisition needs of the economic agents were searched, words are related to the construction sector (mainly searches related to construction materials such as cement and rod,

purchases in hardware stores, etc.) and the set of capital goods imported by the country that are presented in the list of the General Direction of Customs (DGA). The second source used is a replica of the Asymmetric Hashing algorithm (used by Google Correlates[5]) and the K-Nearest Neighborhood (KNN) algorithm (see Vanderkam *et al.,* 2013, for methodological details), in order to identify the keywords related to a particular variable, since the Google platform Correlates is not enabled for the case of the Dominican Republic. The monthly series of the IMAE was obtained from the website of the Central Bank of the Dominican Republic.

For the MLP model, the independent variables are transactions linked to payment systems (in this case, points of sale operations with credit and debit cards) as well as the weighted average interest rate and the volume of sales registered for each month.

3.2. Methodology

3.2.1. Bayesian Structural Time Series Model (BSTS)

State-space models are representations of dynamic systems that provide sufficient flexibility for modeling and treating a wide range of problems in the framework of time series, simplifying the analysis with a unified methodology, as has been widely documented in the economic literature. (Durbin and Koopman, 1999). The main objective of this type of model is to quantify the uncertainty inherent in an unobserved variable, given some information pertaining to a variable.

It stands to reason that the complexity of the state-space representation generally depends on the size of the application considered, the nature of the model, and individual needs. A wide range of models can be expressed in state-space form, allowing selection of components to model the trend, seasonality, regression modulus, and other state components that may be considered relevant.

A structural model of time series has the particularity that it can be expressed in state-space format. As previously pointed out, under this approach, it is assumed that the evolution of the system through time is determined by a series of unobserved vectors $\alpha_1, \dots, \alpha_n$, which are associated to a set of

[5] Google Correlates was available until January 2020; nevertheless, the actualizations of the dataset, after January 2020, were made using a replica of the Asymmetric Hashing algorithm and the KNN algorithm. Google Correlates considers millions of possible candidates (variables) to identify the time series that keep a higher Pearson correlation with the information provided by the user.

observations $y_1, \dots, y_n$. A simple and generalized state-space representation that reflects the relationship between the vector of y_t and α_t is given by a system of first-order differential equations, which can be expressed as follows:

$$y_t = Z_t\alpha_t + \varepsilon_t,\ \varepsilon_t \sim N(0, H_t), \qquad (1)$$
$$\alpha_{t+1} = T_t\alpha_t + R_t\eta_t,\ \eta_t \sim N(0, Q_t)\ t=1,\dots,n,$$

where y_t is a $p\times1$ vector of observations called the observation vector and α_t is an unobserved $m\times1$ vector called the state vector. The first equation of system (1) is called the observation equation, and the second equation is called the equation of state. The states are assumed to follow a Markovian transition process.

Considering that it is not possible to observe $\boldsymbol{\alpha_t}$ directly, the analysis must be based in the observations y_t. The arrays $Z_t, T_t, R_t, H_t\ y\ Q_t$ are initially assumed to be known and the error terms ε_t, η_t are assumed to be serially and mutually independent in each moment of time t. The arrays Z_t and T_{t-1} can depend on $y_1,\dots, y_{t-1}$. The initial state vector α_1 is expected to be $N\ (\alpha_1, P_1)$ and independent of $\varepsilon_1, \dots, \varepsilon_n$ and $\eta 1,\dots,\eta n$, where it is assumed that, initially, , α_1 and P_1 are known. The first equation in (1) has the structure of a linear regression model, where the vector of coefficients α_t varies over time. The second equation represents an auto-regressive vector (VAR) of first order.

Fort the purpose of this research, the considered state space representation is the one suggested by Scott & Varian (2014):

$$\begin{aligned} y_t &= \mu_t + \tau_t + \beta^T x_t + \varepsilon_t \\ \mu_t &= \mu_{t-1} + \delta_{t-1} + u_t \qquad (2) \\ \delta_t &= \delta_{t-1} + v_t \\ \tau_t &= -\sum_{s=1}^{S-1} \tau_{t-s} + \omega_t \end{aligned}$$

where it is pointed that $\eta_t = (u_t, v_t, \omega_t)$ contains independent components of gaussian noise. In this case, Q_t is a constant diagonal matrix, with diagonal elements $\sigma_\varepsilon^2, \sigma_v^2\ y\ \sigma_\omega^2$; H_t is a constant. The advantage of this model relies on the fact that it captures both the trend and seasonal component

and, additionally it incorporates exogenous information. The level of the trend in time t is represented μ_t and the slope is δ_t.

The seasonal component, τ_t, can be represented as a set S of dummy variables with dynamic coefficients that are adjusted to the restriction of having an expected value equivalent to zero during the complete cycle of S seasons. The parameters in (2) are the variances $\sigma_\varepsilon^2, \sigma_u^2, \sigma_v^2, \sigma_\varpi^2$ and the vector of coefficients of the regression component β.

In addition to the apparent complexity involved in the implementation of the proposed modeling technique, the methodologies used to treat this type of model are relatively simple, with the Kalman filter, the Kalman smoother and the bayesian data augmentation being the most popular techniques. Nevertheless, it is necessary to point out that both the Kalman filter and the Kalman smoother provide analytical solutions and are appropriate for the estimation of Gaussian models, in which normality and linearity are assumed, while non-Gaussian models require a more extensive treatment. advanced, such as bayesian data augmentation or other smoothing techniques.

The Kalman filter (Kalman, 1960) is a recursive algorithm that incorporates the information provided to process all the available measurements, with the purpose of estimating the current value of the variables of interest (Maybeck, 1979). It is said to be recursive since it does not require all previous information to be saved, stored, and reprocessed each time a new measure is entered. Denoting $y_t = y_1, \dots, y_t$ as a vector that represents the dependent variable, the Kalman filter computes recursively the predictive distribution $p(\alpha_{t+1}|y_{1:t})$, combining $p(\alpha_t|y_{t-1})$ with y_t and using an standard set of formulas equivalent to a linear regression.

On the other hand, the Kalman smoother moves backward through time, distributing information about the latest observations to previous pairs (μ_t, P_t), where μ_t represents the mean of the distribution and P_t is the variance. The Kalman smoother returns the T means and variances of the unobservable variables, at each moment of time within the sample and conditioned to the information of the complete sample.

It is evident that filtering and smoothing are the most used computational techniques to deal with state-space models. However, when the assumptions of normality and linearity are not plausible, it is not possible to generate exact calculations based on the Kalman filter and the Kalman smoother;

these methods are not pragmatic in the treatment of high-dimensional problems (Pnevmatikakis *et al.,* 2012). In this case, it is necessary to appeal to other methodologies that contemplate non-Gaussian assumptions. In this sense, various approaches have been proposed (Fahrmeir & Wagenpfeil, 1997; Fahrmeir & Tutz, 1994, Jungbacker & Koopman, 2007), including a Bayesian smoother that employs Markov chains using Monte Carlo[6] simulations.

The computation based in Bayesian data augmentation (Neal & Kypraios, 2015; Taylor *et al.*, 2015) produces simulations from $\boldsymbol{p(\alpha_t|y)}$, where $\boldsymbol{y}$=$y_{1:n}$ y $\boldsymbol{\alpha}$=$\alpha_{1:n}$ denote the complete sets of observe and latent data. This alternative is presented given that it is not possible to derive each $\boldsymbol{\alpha_t}$ from $\boldsymbol{p(\alpha_t|y)}$, considering that the serial correlation between α_t y α_{t+1} must be considered.

The Bayesian approach is inherent in a theory where the state parameters are random variables and the observations are fixed; for a Bayesian treatment, simulation-based methods are needed for the estimation of additional parameters. One of these simulation-based methods is the Durbin and Koopman (2002) algorithm, which is used in this case (Scott and Varian, 2014). The advantage of this approach is that it extends the scope of the Kalman smoother, making it possible to answer questions concerning the covariances of unobservable variables over time (which is a limitation of the Kalman smoother). The Durbin and Koopman simulation smoother are among the fastest and most convenient to implement in computational terms and is briefly synthesized by Jarocinski (2015) as follows:

Step 1: Derive α^+ and y^+ through the recursion in (1), where this recursion is initialized $\alpha_1^+ \sim N(0, P_1)$.

Step 2: Generate an artificial serie $y^* = y - y^+$ and derive $\widehat{\alpha^*}$=$E(\alpha|y^*)$ passing y^* through the Kalman filter and the Kalman smoother.

Step 3: Generate $\hat{\alpha} = \widehat{\alpha^*} + \alpha^+$, where $\hat{\alpha}$ is derived from the distribution of α conditional on y.

The state-space representation of the system (2), considered to carry out the estimates in this research, contains a regression component that allows the inclusion of a set of exogenous factors with the potential to contribute to the minimization of forecast errors; to inquire about how to structure

[6] This is a numerical method that allows to solve problems through the simulation of random variables.

of the model matrices incorporate the regression component in the state-space representation, see Scott & Varian (2014).

The approach based on "spike and slab" priors is presented as a logical option for the estimation of the regression coefficients, since it minimizes the dispersion problem, reducing the size of the regression problem. This prior assumes that the regression coefficients are mutually independent with a bimodal distribution consisting of a uniform distribution (slab) and a distribution that degenerates in the neighborhood of zero (spike). The fundamental idea is to make non-significant coefficients within the data set considered zero, making their posterior mean values small enough. This property of the selection process based on selection through spike & slab[7] priors makes this approach attractive and popular in the bayesian paradigm.

Briefly, given a subset $\beta\gamma$ of elements and let $\gamma_k = 1$ if $\beta_k \neq 0$ and $\gamma_k = 0$ if β_k=0. Mathematically, a spike and slab prior can be express as follows:

$$p(\beta, \gamma, \sigma_\varepsilon^2) = p(\beta_\gamma|\gamma,\sigma_\varepsilon^2)\, p(\gamma) \quad (3)$$

The marginal distribution $p(\gamma)$ in (3) is the spike, as it assigns a higher probability in the epsilon neighborhood of zero[8].). In pragmatic terms, it is convenient to simplify (3) using an independent Bernoulli prior (Scott & Varian, 2014):

$$\gamma \sim \prod_{k=1}^{K} \pi_k^{\gamma k}\ (1-\pi_k)^{1-\gamma k} \quad (4)$$

The equation (4) can be reduce assuming that every ϕ_k has the same value ϕ. The value ϕ can be determined through the expert criteria concerning the expected size of the model. Hence, if an amount p of predictors different than zero is expected, then ϕ=p/K where K is the dimension of ϕ_t. On the other hand, the equation (5) is called the slab, as it is feasible to select the priors that make it weakly informative (converging to a plane surface), conditional on γ:

[7] For more mathematical details about the spike and slab priors, see Mitchel & Beauchamp (1988) or Ishwaran *et al.* (2010).
[8] As in the case of Delta de Dirac.

$$\beta_\gamma|\sigma_\varepsilon^2,\gamma\sim N\left(b_\gamma,\sigma_\varepsilon^2\left(\Omega_\gamma^{-1}\right)^{-1}\right)\quad \frac{1}{\sigma_\varepsilon^2}|\,\gamma\sim Ga(\frac{v}{2},\frac{ss}{2})\quad (5),$$

where $Ga(r,s)$ denotes the gamma distribution with mean r/s and variance r/s^2. The parameter of higher dimension in the equation (5) is the matrix of complete information for the priors of the model Ω^{-1}.

In order to estimate the posterior distribution, a Monte Carlo Markov Chain (MCMC) algorithm is used. Basically, the underlying idea of the MCMC is to evaluate the posterior mean of $x(\alpha)$ for the vector of parameters $\boldsymbol{\phi}$=(θ,β,σ_ε^2,α) through simulations, selecting samples of an augmented joint density $p(\phi,\alpha|\boldsymbol{Y})$. Under this scheme, the sampling from the joint density is generate as a Markov[9] chain.

After the initialization for $\boldsymbol{\phi}$, say $\boldsymbol{\phi}$=$\phi(0)$, in the cycle of the subsequent simulations, the following steps are repeated:

1. Simulate the latent state $\boldsymbol{\alpha}$ from $p(\alpha|y,\theta,\beta,\sigma_\varepsilon^2)$.

2. Simulate $\theta\sim p(\theta|y,\alpha,\beta,\sigma_\varepsilon^2)$.

3. Simulate β y σ_ε^2 from a MCMC with stationary distribution $p(\beta,\sigma_\varepsilon^2\ |\boldsymbol{y},\boldsymbol{\alpha},\theta)$.

A sequence $\phi^{(1)},\phi^{(2)},\dots,\phi^{(n)}$ is generated from a Markov chain with stationary distribution $p(\phi|\boldsymbol{y})$, the posterior distribution of ϕ given y. Under the Bayesian paradigm, it is pertinent to generate simulations from the predictive posterior distribution (Scott and Varian, 2014), given the draws of the model parameters and the state of the predictive distribution. The iterations can be performed directly from the system of equations given by the specification (2), starting from the specification $a_n^{(g)}$ with parameters $\theta^{(g)}$, $\beta^{(g)}$, $\sigma_\varepsilon^{2(g)}$, with the purpose of generating a sample of $p(\tilde{y}|\boldsymbol{y})$ through $p(\tilde{y}^{(g)}|\phi^{(g)})$ and, given a set of draws from Monte Carlo simulation, the predictive distribution draws help to automatically correct irrelevant coefficients. This projection method generates a sample of draws from the predictive distribution that can be summarized by their respective means, which is a Monte Carlo estimate of $E(\tilde{y}|\boldsymbol{y})$.

[9] In probability theory, a special type of discrete stochastic process in which the probability of an event occurring depends on the immediately preceding event is known as a Markov chain.

3.2.2. Neural Networks

The use of neural networks to model the behavior of macroeconomic variables has become increasingly popular in the empirical literature (Jalil & Missas, 2007; Tkacz & Hu, 1999; Ballı & Tarimer, 2013), given their ability to specify models with minimum forecast errors, as well as to capture patterns or non-linear components. In this order, it is considered that the appropriate specification of these structures provides greater precision in the results derived from non-linear functions, thus positioning themselves as important tools in the preparation of forecasts of macroeconomic and financial variables, with great predictive capacity.

A neural network can be defined as a structure of simple processing elements, called nodes, whose learning ability is stored in the strength of the connection units or weights, obtained from a training process or learning phase of a set of patterns (Hayking, 2008). The architecture of a neural network tries to emulate the design and functioning of the human brain, and the way in which the neurons of the network are organized is closely linked to its training algorithm; if the network has learned the underlying structure of the problem at hand, then it should be able to classify and predict subsequent patterns (Gurney, 1997).

Kuan and Liu (1995) point out that neural networks can be visualized as input-output models, which can be understood as nonlinear regression functions that characterize the relationship between a dependent variable (input) Y_t and a vector of explanatory variables (output) $X_t = (x_{1t}, \dots, x_{pt})$. Without considering a specific nonlinear function, the model is built by combining nonlinear functions through compositions ranging from single layer neurons to multilayer structures. A neural network structure given by the following expression is used:

$$Y_t = \sum_{j=1}^{j} X_j \alpha_j \quad \alpha_j > \alpha_j' \qquad (6)$$

where Y_t represents the output variable (output), in this case the observed variable for a given period, and X_j represents the network input. On the other hand, α_j is a hyperparameter of the network, which determines the activation of the neuron to transmit the information. This activation is usually represented by a logistic function (in this case, a sigmoid function) given by:

$$f(\mu) = \frac{1}{1+e^{-\mu}} \qquad (7)$$

Since only one hidden layer is used in the model, H_j, equation (6) can be rewritten as follows:

$$H = f\left[\sum_{j=0}^{j} x_j \alpha_j\right] \quad (8)$$

Denoting θ_j as the weight that links the input and the output of model, then:

$$Y_t = \sum_{j=0}^{j} H_j \theta_j \quad (9)$$

Replacing (8) in (9), the function of *h* hidden layers is obtained, including the input function g:

$$Y_t = h\left[\left(\sum_{k=1}^{k} \alpha_k\right) f\left(\sum_{j=0}^{j} \theta_{ik} X_j\right)\right], \quad (10)$$

where:

j=an input network with one layer;

k= two hidden neurons with one layer;

The literature does not expose a definitive rule to carry out the optimal selection of hidden layers and neurons. However, it is pointed out that one of the strategies to determine these hyperparameters is linked to the performance of the model in the testing phase, observing that there is not a saturated neural network, which would lead to an overfitting or, on the contrary, to an underfit. Similarly, one of the essential rules (Hornik, 1991) is that the number of neurons must be bounded by the number of inputs and outputs of the model. Regarding the number of hidden layers, Cybenko's theorem (1989) is followed, which states that, with a hidden layer and a finite number of neurons, it is possible to approximate continuous functions with assumptions about the activation function (in this case, as previously pointed out, a sigmoid activation function is used).

Once the number of hidden layers and neurons has been established, we seek to minimize the function:

$$min_{\alpha k \theta(j,k)} SSD = \sum_{t=1}^{T}\left[Y_t - h\left(\sum_{k=1}^{k} \alpha_k f\left(\sum_{j=0}^{j} \theta_{ik} X_{jt}\right)\right)\right]^2 \quad (11)$$

It is important to note that, in addition to the selection of neurons and layers of the model, this minimization process requires a training sample, in which the network "learns" from the data

provided as inputs, in order to carry out the phase of predictions; in this case, the training sample represents 60% of the total data, as well as a "test sample", in order to verify the accuracy of the predictions achieved by means of the trained sample (Tkacz & Hu, 1999 ; Basihos, 2016).

3.2.3. Lasso & Ridge Regressions

Lasso and ridge regressions are variants of the linear regression model, where regularization terms are incorporated in order to reduce overfitting problems. In the case of ridge regression (Tikhonov regularization), a shrinking process is carried out on all the coefficients in an epsilon neighborhood of zero (ε>0); under this approach, the regularization term is equivalent to $\alpha \sum_{i=1}^{n} \theta_i^2$, which is added to the cost function (Géron, 2017). Through the hyperparameter α, the degree of regularization of the model is adjusted:

$$J(\theta) = MSE\ (\theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2 \qquad (12)$$

If $\boldsymbol{w}$ is defined as the vector of the feature weights $(\theta_1\ a\ \theta_n)$[10], then the regularization term is equivalent to $\frac{1}{2}(\|w\|_2)^2$, where $\|\ .\ \|_2$ represents the ℓ_2 norm of the weights of the vectors. On the other hand, lasso regressions (Least Absolute Shrinkage and Selection Operator Regression), like ridge regressions, impute a penalty term in the cost function; however, in this case, the ℓ_1 norm of the weighting vector is used instead of the ℓ_2 norm. The lasso regression cost function is:

$$J(\theta) = MSE(\theta) + \alpha \sum_{i=1}^{n} [\theta_i] \qquad (13)$$

In lasso regressions, the variables that do not have a significant contribution to explain the behavior of the dependent variable are completely eliminated.

IV. Results

4.1. BSTS model

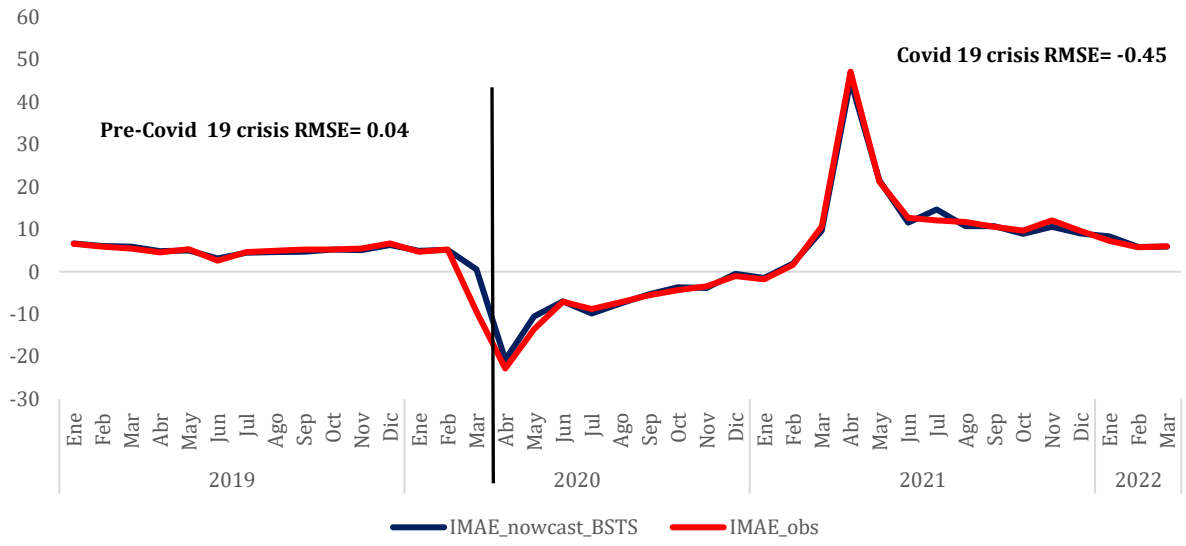
The nowcasting exercises in the frame of this document, both in and out of sample, are carried out using the methodology described in the section III. In the case of the BSTS model, the “spike and slab” priors are considered with the purpose of debugging the variables that are not significant and

[10] The sum is initialized *i=1, given that the term* θ_0 is not regulated.

minimize the dispersion problem. The time interval considered for the estimation for the IMAE is january 2017- march 2022, using 276 variables (Table I.3, Appendix I). After selecting the explanatory variables for the BSTS model, the first stage of the estimation process consisted in mixing the frequency of the data, given that the data for the independent variables is in a different frequency than the data for the dependent variables.

In a second phase, the specification and estimation for the time series component is carried out, considering a linear trend and a seasonal component, as well as the inclusion of the regression vector. In the graphs of Appendix I, it is shown the distribution for each one of these components. It is possible to infer useful information about the prior, considering the expected dimension for the matrix of the explanatory variables, the expected R^2, the size of the sample v and using diagonal shrinkage[11], that corresponds to the prior g of Zellner. The other parameters were established with the values that the BSTS package (Scott, 2016) gives by default (i.e. R^2; v=0.01); the amount of iterations for each case was set in 5,000.

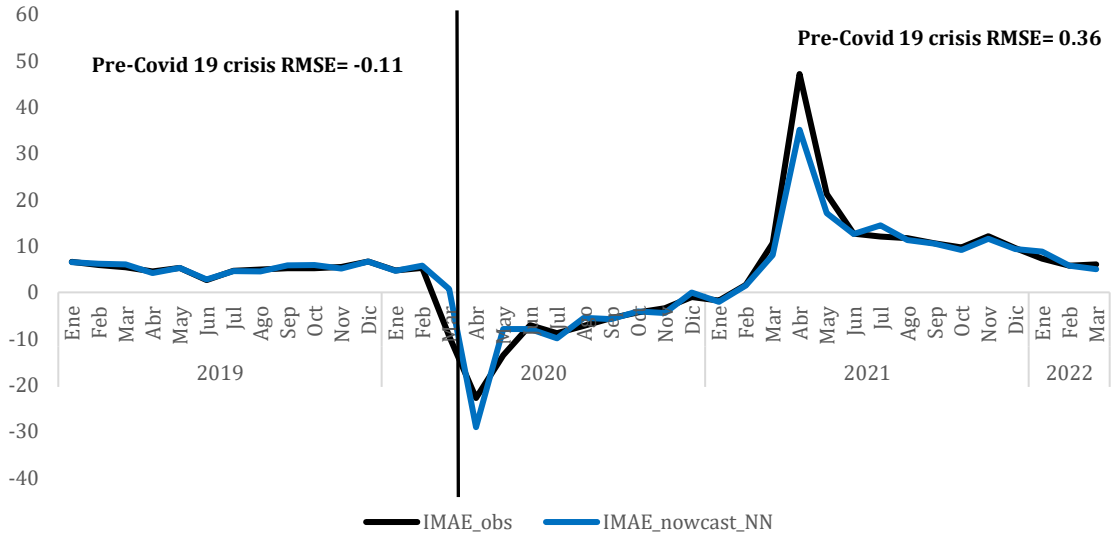
Figure 4.1. One-step ahead nowcast for the monthly indicator of economic activity (IMAE), year-over-year variation (%), BSTS model.



**Source: author's elaboration with the BSTS model.*

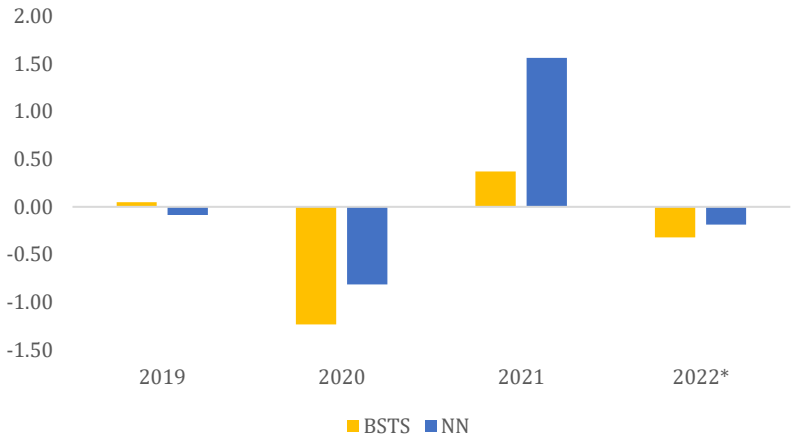
[11] For additional information about the diagonal shrinkage, see Scott (2016).

Figure 4.2. One-step ahead nowcast for the monthly indicator of economic activity (IMAE), year-over-year variation (%), MLP model.



Source: authors elaboration with MLP model.

Figure 4.2.a. Root Mean Square Error (RMSE) for MLP and BSTS model.



Source: authors' elaboration.

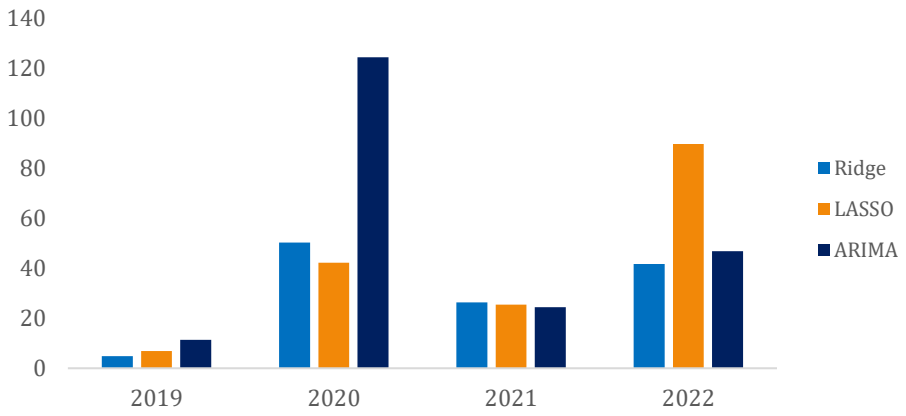
4.2. Ridge & LASSO

The Ridge and LASSO regressions where performed using a rolling window sample of 5 years. This means that for every step of forecast, data from the previous 260 weeks (last 5 years) is used. Both models use information from Google Trends (272 variables related to consumption and investment which are closely linked to economic activity) and other variables (4 variables related to gas prices and exchange rate). The full list of variables is shown in the Appendix IA. For every step of the forecast process a cross-validation

For every step of the forecast process a cross-validation [12]is performed using 10 subsets of the sample. In order to determine the optimal value of the hyperparameter α, which minimizes the cost function defined in 3.2.3, a vector of 100 possible values are tested. After the cross-validation process ends testing each the 100 α the one that minimizes the cost function is selected and used to forecast

The mean square error (MSE) is used as measure of performance, shown on Figure 4.3. On average, both models (Ridge and LASSO) had better performance in 2019 than the following years (2020 - 2022M03). The increase on average MSE is caused by the COVID-19 pandemic shock. In general terms, Ridge and LASSO regressions show lower MSE than the benchmark model (auto ARIMA).

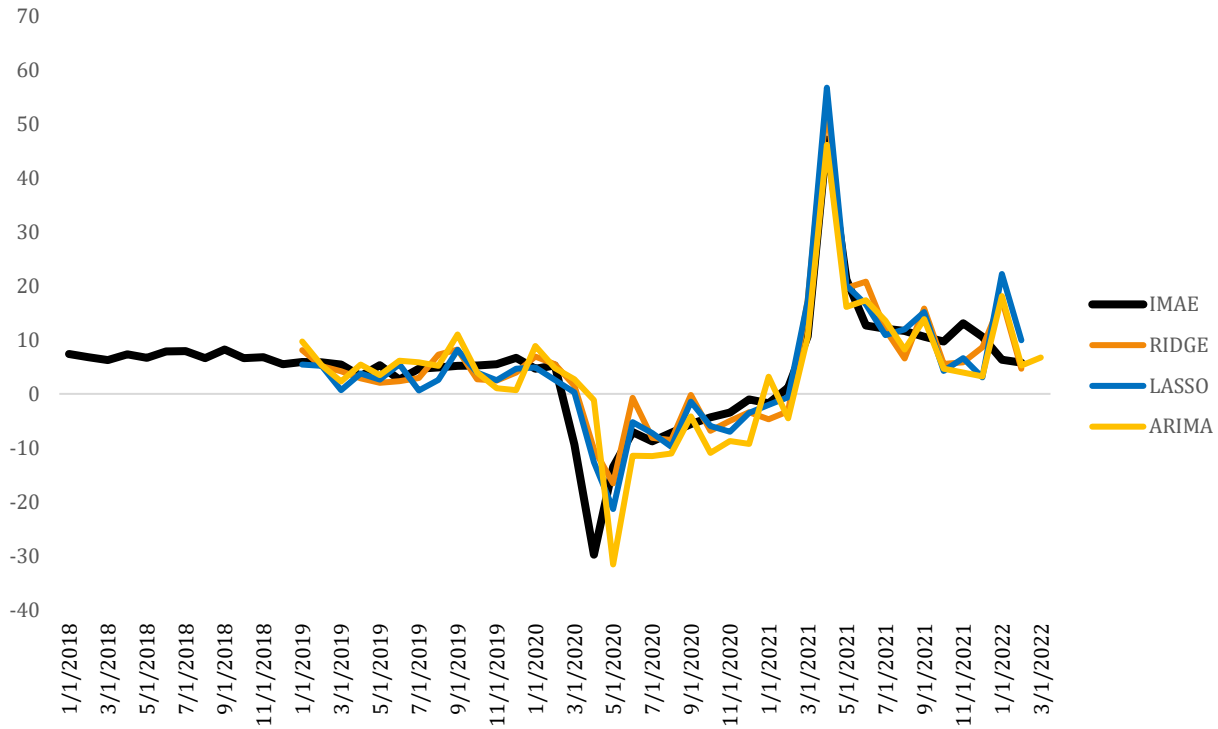
Figure 4.3: Average mean square error of one-step ahead forecast.



**Source: authors' elaboration; results of lasso and ridge regressions.*

[12] Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

Figure 4.4: One-step forecast of monthly economic activity.



**Source: authors' elaboration; results of lasso and ridge regressions.*

Appendix IA. List of independent variables for the BSTS model and for the Lasso and Ridge regressions.

Category	Term	Location
All categories	capital goods + exports	US
All categories	exports + dominican republic	US
All categories	exports + dominican republic	JP
All categories	exports + dominican republic	CN
All categories	prestamos + personales	DO
All categories	accesorios + hogar	DO
All categories	precios + bebidas alcoholicas	DO
All categories	tipo de cambio	DO
All categories	tipo de cambio + compra	DO
All categories	tipo de cambio + venta	DO
All categories	prestamos personales	DO
All categories	electricistas + servicios	DO
All categories	doncella + servicios	DO
All categories	agencia + "servicios domesticos"	DO
All categories	plomeros + servicios	DO
All categories	millas + "banco popular"	DO
All categories	banco bhd	DO
All categories	banreservas	DO
All categories	courier + servicios	DO
All categories	vimenca	DO
All categories	moneygram + remesas	DO
All categories	aeropaq	DO
All categories	tasa de interes	DO
All categories	precios	DO
All categories	dollar	DO
All categories	gold + gym	DO
All categories	taxi	DO
All categories	precios + alimentos	DO
All categories	precios + viviendas	DO
All categories	ferreteria	DO
All categories	paypal	DO
All categories	Payoneer	DO
All categories	precios + casa	DO
All categories	precio + apartamentos	DO
All categories	precio + solares	DO
All categories	costo	DO
All categories	restaurant	DO
All categories	uber	DO

Category	Term	Location
All categories	"internet banking"	DO
All categories	reservar + hotel	DO
All categories	cine + tickets	DO
All categories	caribbean + cinema	DO
All categories	palacio del cine + tickets	DO
All categories	"cartelera cine"	DO
All categories	boleteria + "teatro nacional"	DO
All categories	ticket express	DO
All categories	automovil	DO
All categories	precios + automoviles	DO
All categories	feria + vehiculos	DO
All categories	venta + vehiculos	DO
All categories	carro + toyota	DO
All categories	carro + honda	DO
All categories	hyundai + carro	DO
All categories	mazda	DO
All categories	supercarro	DO
All categories	kia	DO
All categories	auto + parts	DO
All categories	seguros + autos	DO
All categories	salon de belleza + spa	DO
All categories	spa	DO
All categories	maquillaje + comprar	DO
All categories	beauty + supply	DO
All categories	clinica + estetica	DO
All categories	gym	DO
All categories	computadoras + precios	DO
All categories	electrodomesticos + precios	DO
All categories	muebles + hogar	DO
All categories	servicios + reparaciones	DO
All categories	iphone + comprar	DO
All categories	piezas + computadoras	DO
All categories	piezas + celulares	DO
All categories	servicios + internet	DO
All categories	ferreterias + telefonos	DO
All categories	materiales + construccion	DO
All categories	compra + cemento	DO
All categories	construccion + viviendas	DO
All categories	costo metro cuadrado + construccion	DO
All categories	precio + mano de obra + construccion	DO
All categories	precios materiales + construccion	DO

Category	Term	Location
All categories	construccion	DO
All categories	constructoras + republica dominicana	DO
All categories	precios + blocks	DO
All categories	sector + construccion	DO
All categories	varillas + precios	DO
All categories	ramon corripio + ferreteria	DO
All categories	ferreterias + republica dominicana	DO
All categories	ferreteria + maderera central	DO
All categories	ferreteria + popular	DO
All categories	ferreteria + cima	DO
All categories	ferreteria + hermanos pappaterra	DO
All categories	caribe group + santo domingo	DO
All categories	ferreteria + felimon	DO
All categories	ferreteria americana	DO
All categories	ferreteria ochoa	DO
All categories	salarios + construccion	DO
All categories	compra + pino	DO
All categories	precios + madera	DO
All categories	importadora + madera	DO
All categories	precios + plywood + republica dominicana	DO
All categories	precios + hormigon	DO
All categories	precios + arena	DO
All categories	precios + grava	DO
All categories	precios + pintura	DO
All categories	compra + sanitarios	DO
All categories	supermercado nacional	DO
All categories	tasa del dolar	DO
All categories	compra + pisos	DO
All categories	maquinas + empaquetadoras	DO
All categories	aparatos + mecanicos	DO
All categories	precios + tractores	DO
All categories	yeso + precios	DO
All categories	maquina electrica + precios	DO
All categories	equipos + electricos	DO
All categories	precios + niquel	DO
All categories	precios + hierro	DO
All categories	precios + cobre	DO
All categories	precios + aluminio	DO
All categories	plomo + precios	DO
All categories	zinc + precios	DO
All categories	estano + precios	DO

Category	Term	Location
All categories	calderas + precios	DO
All categories	calderas + compra	DO
All categories	tiendas + armas online usa	DO
All categories	maquinas + mecanicas + compras	DO
All categories	maquinas electricas + compras	DO
All categories	material electrico + precios	DO
All categories	repuestos + maquinarias	DO
All categories	electricos	DO
All categories	caterpillar tractores + precios	DO
All categories	john deere + precios	DO
All categories	john deere + catalogo	DO
All categories	ferreteria americana + precios	DO
All categories	energia + costos	DO
All categories	pago online + edeeste	DO
All categories	gasolina + precios	DO
All categories	combustibles + precios	DO
All categories	claro + "pago online"	DO
All categories	edeeste + "pago online"	DO
All categories	edenorte + "pago online"	DO
All categories	edesur + "pago online"	DO
All categories	windtelecom + servicio	DO
All categories	caasd + pago	DO
All categories	caasd	DO
All categories	ofertas + supermercados	DO
All categories	precios + supermercados	DO
All categories	precios + arroz	DO
All categories	precio + canasta basica	DO
All categories	precios + cebolla	DO
All categories	precios + carne	DO
All categories	precios + aceite	DO
All categories	precios + habichuelas	DO
All categories	precio + pollo	DO
All categories	precios + viveres	DO
All categories	precio + platano	DO
All categories	jumbo	DO
All categories	la sirena	DO
All categories	supermercado bravo	DO
All categories	supermercado + pola	DO
All categories	farmacia	DO
All categories	farmax	DO
All categories	farmacia gbc	DO

Category	Term	Location
All categories	farmacia value	DO
All categories	farmacia carol	DO
All categories	laboratorio referencia	DO
All categories	laboratorio amadita	DO
All categories	plaza de la salud	DO
All categories	cedimat	DO
All categories	centro + ginecologia y obstetricia	DO
All categories	clinica + san rafael	DO
All categories	centro medico dominicano	DO
All categories	clinica abreu	DO
All categories	instrumentos + medicos	DO
All categories	comprar + equipos medicos	DO
All categories	instrumentos opticos + comprar	DO
All categories	instrumentos opticos + importar	DO
All categories	equipos + medicos	DO
All categories	medical + equipment + import	DO
All categories	equipos medicos + precios	DO
All categories	instrumentos opticos + precios	DO
All categories	instrumentos oftalmologicos + importar	DO
All categories	aparatos quirurgicos + precios	DO
All categories	seguros pepin	DO
All categories	la colonial + seguros	DO
All categories	mapfre	DO
All categories	seguros universal	DO
All categories	senasa	DO
All categories	ars + "plan salud"	DO
All categories	ars humano	DO
All categories	ars palic	DO
All categories	seguros + vehiculos	DO
All categories	ikea	DO
All categories	decoraciones	DO
All categories	ilumel	DO
All categories	casa cuesta	DO
All categories	complementos + hogar	DO
All categories	agencia + "bienes raices"	DO
All categories	remax	DO
All categories	supercasas	DO
All categories	retail + price	DO
All categories	precio + minorista	DO
All categories	blusas + damas	DO
All categories	pantalones + damas	DO

Category	Term	Location
All categories	"ropa interior" + damas	DO
All categories	victoria secret	DO
All categories	victorias secret + shop	DO
All categories	zara	DO
All categories	pantalones + hombres	DO
All categories	camisas + hombres	DO
All categories	ropa interior + hombres	DO
All categories	old navy	DO
All categories	gap + "shop online"	DO
All categories	zapatos + hombres	DO
All categories	zapatos + mujeres	DO
All categories	women + shoes	DO
All categories	women + clothes	DO
All categories	women + lingerie	DO
All categories	men + clothes	DO
All categories	men + shoes	DO
All categories	forever21 + shop	DO
All categories	ebay	DO
All categories	celulares + venta	DO
All categories	online + shop	DO
All categories	amazon	DO
All categories	reservas + vuelos	DO
All categories	cheap + flights	DO
All categories	resort	DO
All categories	corotos.com	DO
All categories	la casa del colt	DO
	Advertising & Marketing	DO
	Agricultural Equipment	DO
	Agriculture & Forestry	DO
	Alcoholic Beverages	DO
	Automotive Industry	DO
	Autos & Vehicles	DO
Banking		DO
	Bicycles & Accessories	DO
	Campers & RVs	DO
	Candy & Sweets	DO
	Classic Vehicles	DO
	Commercial Vehicles	DO
	Computer Hardware	DO
	Construction & Maintenance	DO
	Consulting	DO

Category	Term	Location
	Cosmetic Procedures	DO
	Cosmetology & Beauty Professionals	DO
	Credit & Lending	DO
	Currencies & Foreign Exchange	DO
	E-Commerce Services	DO
	Electronics & Electrical	DO
	Energy & Utilities	DO
	Enterprise Resource Planning (ERP)	DO
Fitness		DO
	Food Production	DO
	Grocery & Food Retailers	DO
Hair Care		DO
	Hospitality Industry	DO
	Hybrid & Alternative Vehicles	DO
Hyundai		DO
	Insurance	DO
Investing		DO
	Make-Up & Cosmetics	DO
Mazda		DO
	Motorcycles	DO
	Office Services	DO
	Office Supplies	DO
	Pharmaceuticals & Biotech	DO
	Skin & Nail Care	DO
Software		DO
	Spas & Beauty Services	DO
Toyota		DO
	Unwanted Body & Facial Hair Removal	DO
	Weight Loss	DO