Machine Learning Using Auto Regressive Vectors to Predict Ecuador's Percentage Growth

Marcelo León[1](✉), Fabricio Echeverria[2], Veronica Huacon[3], and Darlys Sares[1]

[1] Universidad ECOTEC, Samborondon, Ecuador
marceloleon11@hotmail.com
[2] Corporación Universitaria de Asturias, Bogota, Colombia
pedro.echeverria@asturias.edu.co
[3] Universidad de Guayaquil, Guayaquil, Ecuador

Abstract. Automated machine learning, known as AutoML, is a process that utilizes automated techniques and algorithms to develop machine learning models. This encompasses feature selection and preprocessing, algorithm selection, hyperparameter optimization, and model performance evaluation. The goal of AutoML is to simplify and expedite the model building process, enabling individuals with less data science experience to construct effective models without the need to manually navigate all the stages.

In the context of an economic study on Ecuador, it is proposed to employ AutoML to analyze the country's economic data and forecast the Gross Domestic Product (GDP) growth. Various factors that have historically impacted Ecuador's GDP are taken into consideration, such as reliance on the export of a single raw material, political and economic instability, fiscal deficits, fluctuations in oil prices, financial crises, and the COVID-19 pandemic.

It is suggested to utilize a Vector Autoregressive (VAR) model within the machine learning process. VAR is a statistical model that allows us to comprehend how various variables co-vary over time. Additionally, different regression analysis algorithms are mentioned, such as decision trees, k-nearest neighbors, random forest, support vector machines, and artificial neural networks, which can be employed in the analysis of economic data.

Keywords: AUTOML · Vector Autoregressive · GDP · Supervised Learning

1 Introduction

For a considerable period, mathematical techniques like time series analysis have been employed for predicting the behavior of a variable using modeling methods such as ARMA, ARIMA, or SARIMA. However, these techniques lacked the capability to incorporate new variables of various types, such as binary representations, images, or other forms of data, which are currently extensively utilized.

The advent of machine learning has facilitated the integration of diverse data types, and models can be constructed through Regression Analysis employing algorithms like

A. Bajaj et al. (Eds.): SoCPaR 2023, LNNS 1247, pp. 192–200, 2025.
https://doi.org/10.1007/978-3-031-81086-2_21

Regression Trees, Random Forest, Support Vector Regression, k-Nearest Neighbors, or Artificial Neural Networks (with only one layer of hidden neurons).

AutoML (Automated Machine Learning) pertains to the automation of the machine learning model development process. It entails the utilization of automated techniques and algorithms to perform tasks that traditionally necessitate human intervention, including feature selection, preprocessing, algorithm selection, hyperparameter tuning, and model performance evaluation.

The aim of AutoML is to simplify and streamline the process of constructing machine learning models, enabling individuals with limited experience in data science to leverage these tools and acquire effective models without having to manually navigate all stages of the process. AutoML seeks to reduce technical complexity and expedite model development by harnessing automated algorithms and techniques.

Some of the tasks that can be automated with AutoML include the automatic selection of relevant features, hyperparameter optimization for enhancing model performance, automated detection of data issues such as missing values or noise, and automatic code generation for model implementation and deployment in various environments.

There are two types of questions arising from the analysis of a machine learning application on economic data from Ecuador:

– The first type of question pertains to Machine Learning: What coding methods can be utilized to represent a time series? How can additional variables be encoded and introduced into the time series? What forms of supervised learning can be employed?

  The second type of questions relates to economic data: How is a dataset structured, with identification of time series and additional variables? Where can the different variables that characterize Ecuador's GDP growth be sourced?

2 Literature

The Gross Domestic Product (GDP) serves as a measure of a country's economic production, and its evolution over time reflects the economic development of a nation. Ecuador has faced various factors that have influenced its GDP throughout its history. Some of the most significant factors that have impacted Ecuador's GDP since the beginning of the 20th century include:

Dependence on the export of a single commodity: For a significant portion of the 20th century, Ecuador has relied heavily on the export of a single commodity, such as bananas or oil. This reliance has made its economy vulnerable to fluctuations in the prices of these products in the international market.

Political and economic instability: Throughout its history, Ecuador has experienced significant political and economic instability, marked by numerous changes in government and financial crises. This instability has had a detrimental effect on investor confidence and has contributed to GDP volatility.

Fiscal deficit: Ecuador has grappled with a fiscal deficit for much of its history, leading to excessive indebtedness and inflationary pressures. These factors have had a negative impact on the country's GDP and overall economic growth.

Fluctuations in oil prices: Ecuador is a major exporter of oil, and fluctuations in oil prices have exerted a significant influence on its economy. Low oil prices, in particular, have adversely affected Ecuador's GDP and have contributed to the country's economic instability.

Financial crises: Ecuador has experienced various financial crises throughout its history, including the financial crisis of 1999 and the recessionary crisis of 2015. These crises have had a negative impact on GDP and have contributed to greater economic instability.

COVID-19: Ecuador was also impacted by the coronavirus pandemic (2019-nCoV), later classified as SARS-CoV-2, which causes the disease COVID-19. The pandemic was declared by the World Health Organization on March 11, 2020.

It is important to note that these are just some of the factors that have affected Ecuador's GDP throughout its history, and there are many other factors that have also influenced the country's economic development.

A historical review of the data that have influenced the situation in Ecuador encompasses various aspects affecting the percentage of Ecuador's gross domestic product between 1961 and 2020.
Main time series:

– Percentage growth of Ecuador's gross domestic product (The World Bank, 2021)

Types of Political Events:

– Years of Political Crisis in Ecuador (Arboleda, 2015)
– Presidential election years
– Years of military dictatorship (North, 2006)

Types of Economic Events:

– Inflation (Consumer Prices) (The World Bank, 2021)
– Currency exchange (Dollarization) (Larrea M., 2004)
– Oil Exports (oil price) (Statista, 2021)

Demographic Event Types:

– Population of Ecuador (The World Bank, 2020)
– Total Volume of Migrants (the data presents totalized values every 5 years, which were imputed using a linear interpolation method for the intermediate years) (The World Bank, 2020)

Types of Environmental Events:

– El Niño Phenomenon

Types of Health Events:

– COVID-19

2.1 Vector Autoregressive Model

A Vector Autoregressive (VAR) model is a statistical model used to comprehend how multiple variables change together over time. It falls under the category of multivariate time series models that fit a set of historical data to forecast how these variables will evolve in the future.

To train a VAR model, historical data related to the variables of interest is necessary. These data are organized in a time matrix, with each column representing a variable and each row signifying a specific time point. Once the time matrix is obtained, a linear model is employed to forecast how the variables will evolve in the future based on their past behavior.

A VAR model can be used for predictions in various ways, such as forecasting the future value of a single variable given a set of past values of other variables, or predicting the future values of multiple variables based on their past values. Additionally, it can be employed to understand the relationships between specific variables and how changes in one variable can impact others.

A VAR model proves to be a valuable tool for comprehending how different variables change over time and for making predictions concerning their future behavior.

2.2 Regression Analysis

In the field of computer science, machine learning employs various algorithms for making quantitative predictions, collectively known as Regression Analysis [4]. Among these, we have:

– Decision Trees, which partition the predictor space (independent variables) into distinct, non-overlapping regions. When involving a quantitative independent variable, they are referred to as regression trees [4].
– k-Nearest Neighbors (k-NN), where the choice of distance metric and the assumption that points close to each other are similar are crucial. For quantitative independent variables, it's known as k-NN regression, with the result being the average of neighboring data points [4].
– Random Forest, a model comprised of multiple trees generated from different sets of random data [2]. In the case of regression, it computes the average result from these trees.
– Support Vector Machine (SVM), offering flexibility in defining the acceptable error level for a model and finding an appropriate line (or hyperplane in higher dimensions) to fit the data. When applied to quantitative independent variables, it's known as Support Vector Regression (SVR) [3].
– Artificial Neural Networks, which combine artificial neurons with input and quantitative output variables through activation functions [4] [5].
– C 5.0, an improved algorithm over C 4.5 for regression trees [10] [11].
– Bayesian networks, utilizing probability networks to determine conditions under which a response could be obtained. While primarily designed for classification, it can be adapted for regression by transforming qualitative values into quantitative ones

2.3 Regression Analysis Quality Metrics

Regression analyzes have indicators to measure the results of the model, including:

- Mean Square Error (MSE) is a measure of dispersion that measures the sum of the squared distances between the predicted and actual values of a regression [6].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

- Determination coefficient, also known as R ^{2}, measures the degree of dispersion with respect to the mean of the results and the real values of a regression [6].

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \underline{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \underline{y}\right)^2}$$

- Mean Absolute Error (MAE) is a measure of dispersion that measures the absolute sum of the difference between the predicted and actual values of a regression [6].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

2.4 Features Encoding

These are algorithms that enable the conversion of qualitative data into quantitative data. Domain modeling is kept distinct from characterization to differentiate the more abstract aspects from the more operational ones. Transforming the raw data of a domain into features is inherently a creative approach, akin to modeling in any other realm of mathematics and (software) engineering. Modeling and design, in general, involve a series of decisions and approximations, and as such, they inherently possess an element of imperfection (DUBOUE, 2020).

3 Methodology

Learning with Autoregressive Vectors is defined as follows:
The formulation of a Vector Autoregressive model (VAR) depends on the number of variables to be included in the model and the number of previous periods to consider when making predictions. In general, the formulation of a VAR model can be expressed as follows:

$$\mathrm{X(t) = c + A1X(t-1) + A2X(t-2) + \ldots + An * X(tn) + e(t)}$$

In this formulation, X(t) is the variable to be predicted in period t, c is a constant term, A1, A2,…, An are autoregression coefficients that determine the importance of each previous period in the prediction. ye(t) is the error of the model in period t.

To formulate an Autoregressive (AR) model in machine learning, the following process is followed:

1. Identify the variables you intend to predict and gather historical data related to them.
2. Organize the historical data into a time matrix, where each column represents a variable, and each row corresponds to a specific time point.
3. Determine the order of the AR model. The order of the AR model signifies the number of previous time steps included as predictors in the model. For instance, if the model order is set to 2, then the variable values from the previous two time points will be incorporated as predictors in the model.
4. Specify the number of variables to include in the model. If you aim to predict a single variable, the model will be a one-dimensional AR model. If you are predicting multiple variables, the model becomes a multidimensional AR model.
5. Fit the AR model using a machine learning algorithm to make predictions via regression machine learning.
6. Utilize the trained model to make predictions regarding future variable values.

It's important to note that an AR model is most suitable for making predictions in the near future. If you need to make long-term predictions, you may need to employ a different model, such as a Vector Autoregressive (VAR) model.

To express the prediction model of S_{i+1}, considering all the independent variables:

$$S_{i+1} \sim f(S_i, v_{1i}, v_{2i}, v_{3i}, v_{4i}, v_{5i}, v_{6i}, v_{7i}, v_{8i}, v_{9i}, v_{10i}, v_{11i})$$

When using Feature Encoding, determine

v_{1i}: Percentage of growth of Ecuador's gross domestic product

v_{2i}: Years of Political Crisis in Ecuador

v_{3i}: Presidential election years

v_{4i}: Years of military dictatorship

v_{5i}:: Inflation (Consumer Prices)

v_{6i}:: Currency exchange (Dollarization)

v_{7i}:: Oil Exports (oil price)

v_{8i}: Population of Ecuador

v_{9i}: Total Volume of Migrants, the data presents totalized values every 5 years which were imputed by means of a straight line for the intermediate years

v_{10i}: El Niño Phenomenon

v_{11i}: COVID

v_{ni} is equal to 1, if event i occurred in that period of time. Otherwise it is 0.

i is the Time Period.

Considerations for Supervised Learning.

Number of records = 58

Training Data Percentage = 100%

Evaluation Data Percentage = 100%

Number of training data records = 58

Number of evaluation data records = 58

Training Data Record Rate = [1:58]

Evaluation Data Record Rate = [1:58]

Note: 100% of the data has been considered for Training and Evaluation due to the small number of records.

Python programming

automl library has been used lazypredict

```
# Install the lazypredict package
pip install lazypredict

# Import the LazyClassifier and LazyRegressor classes from lazypredict.Supervised
desde lazypredict.Supervised import LazyClassifier , LazyRegressor

# import the function train_test_split from sklearn.model_selection
desde sklearn.model_selection import train_test_split

# Import the sklearn datasets module
desde sklearn import datasets

# Define the input characteristics (X) using the df2 DataFrame
# Features include: " Child ", " Dollarization ", " Electoral ", " Population ",
# "Migrants", " Average_Price_Oil ", " Oil_Boom ", "COVID", "GDP_1",
# "Child_1", "Dolarization_1", "Electoral_1", "Poblacion_1", "Migrants_1",
# "Average_Oil_Price_1", "Oil_Boom_1" and "COVID_1"
X = df2[[" Child ", " Dollarization ", "Electoral", " Population ", "Migrants",
" Average_Oil_Price ", " Oil_Boom ", "COVID", "GDP_1",
"Child_1", "Dollarization_1", "Electoral_1", "Poblacion_1",
"Migrants_1", "Average_Oil_Price_1", "Oil_Boom_1", "COVID_1"]]

# Fit the model using LazyRegressor
LazyRegressor is used for regression tasks, but LazyClassifier can also be used for
classification
# Assume that the target variables (y) are also defined
models , predictions = LazyRegressor (). fit (X, X, y, y)

# Print the models generated by LazyRegressor
print ( models )
```

lazypredict package is used to apply machine learning techniques. The package is installed and the LazyClassifier and LazyRegressor classes from lazypredict.Supervised are imported .

train_test_split is imported from sklearn.model_selection to split the data into training and test sets.

sklearn datasets module is imported to load a dataset (although it is not used in the provided code) .

Next, the feature matrix X is defined using a DataFrame called df2 and a selection of specific columns.

Finally, the model is fitted using the LazyRegressor, the generated models are stored in the models variable and printed to the screen.

4 Analysis of Results

The result of the quality metrics of the AUTOML models are shown in the following table:

model	Adjusted R- Square	R- Square	RMSE
ExtraTreeRegressor	1.00	1.00	0.00
DecisionTreeRegressor	1.00	1.00	0.00
GaussianProcessRegressor	1.00	1.00	0.00
ExtraTreesRegressor	1.00	1.00	0.00
GradientBoostingRegressor	0.98	0.99	0.37
XGBRegressor	0.97	0.98	0.48
RandomForestRegressor	0.84	0.88	1.16
AdaBoostRegressor	0.75	0.82	1.44
BaggingRegressor	0.73	0.80	1.51
MLPRegressor	0.55	0.68	1.94
TransformedTargetRegressor	0.40	0.58	2.22
LinearRegression	0.40	0.58	2.22
RidgeCV	0.38	0.56	2.27
HuberRegressor	0.38	0.56	2.28
LassoLarsIC	0.37	0.55	2.28
Ridge	0.36	0.54	2.31
SGDRegressor	0.33	0.52	2.36
LinearSVR	0.27	0.48	2.47
Bayesian Ridge	0.19	0.42	2.60
KNeighborsRegressor	0.14	0.38	2.68
HistGradientBoostingRegressor	0.12	0.37	2.71
LGBMRegressor	0.03	0.31	2.84
TweedieRegressor	0.01	0.30	2.87
SVR	-0.08	0.23	2.99
NuSVR	-0.09	0.23	3.01
LassoLarsCV	-0.18	0.16	3.13
ElasticNetCV	-0.18	0.16	3.13
ElasticNet	-0.19	0.15	3.14
LassoCV	-0.19	0.15	3.15
Orthogonal Matching Pursuit	-0.22	0.13	3.18
OrthogonalMatchingPursuitCV	-0.22	0.13	3.18
Lasso	-0.32	0.06	3.31
LarsCV	-0.32	0.06	3.31
DummyRegressor	-0.40	0.00	3.41
LassoLars	-0.40	0.00	3.41
QuantileRegressor	-0.41	-0.00	3.42
PassiveAggressiveRegressor	-0.65	-0.18	3.70
Kernel Ridge	-1.18	-0.55	4.25
RANSACRegressor	-11.52	-7.91	10.19
Lars	-130562.03	-92942.18	1040.99

From what we can appreciate that the models that use algorithms such as: ExtraTreeRegressor, DecisionTreeRegressor, GaussianProcessRegressor and ExtraTreesRegressor, have an RSME equal to zero and R2 equal to 1.

5 Conclusions

Autoregressive Vectors enable the creation of Tidy Data structures, allowing the inclusion of various variables to explore their relationships. This capability facilitates Multivariate Analysis, Causality Analysis, and predictions and projections, making it a powerful tool for economic analysis and other fields that involve interconnected variables.

Autoregressive Vectors can extend time series data, making it compatible with AUTOML and enabling the addition of new variables. This enhancement aids in comprehending diverse economic dynamics and streamlines the decision-making process.

By utilizing AUTOML, we can search for models with the best metrics, simplifying and accelerating the tasks of data scientists. This approach reduces time and effort compared to other manual and traditional methods.

References

Statista. Annual evolution of the average price of crude oil set by the Organization of Petroleum Exporting Countries (OPEC) from 1960 to 2020 (2021). https://es.statista.com/estadisticas/635114/price-medium-of-crude-fixed-by-opec/. Accessed 1 Feb 2021

The World Bank. GDP growth (annual %). https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?locations=EC. Accessed 1 Feb 2021

North, L.: Military and State in Ecuador: military construction and civil dismantlement? Icons **26**(26), 85–95 (2006)

The World Bank. (1 of 2 of 2021). Inflation , consumer prices (annual %) - Ecuador. https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG?locations=EC. Accessed 1 Feb 2021

Larrea, M.C.: Dollarization and human development in ecuador. Icons **19**, 43–53 (2004)

Grove, M.: The Ecuadorian crisis: its economic and social blockades. institute french d'études andes, Quito(2015)

Grus, J.: data science desde Scratch 2nd Edition First Principles with python (Cronin, M., ed.). O'Reilly Media, Inc., Sebastopol (2019)

Orellana Alvear, J.: (12 of 11 of 2018). Decision Trees and Random Forest. https://bookdown.org/content/2031/. Accessed 1 Mar 2021

Sullivan, W.: Machine Learning For Beginners : Algorithms, Decision Tree & Random Forest Introduction. Healthy pragmatic Solutions, New York (2017)

Suykens, J.A.: Regularization, Optimization, Kernels, and Support Vector Machines. CRC Press, New York (2014)

Hernández Orallo, J., Ferri Ramírez, C., Ramírez Quintana, M.J.: Introduction to Data Mining. Pearson Education, Madrid (2004)

Duboue, P.: The Art of Feature Engineering : Essentials for Machine Learning. Cambridge University Press, New York (2020)

Rivera Jorda, E., Raqueno, A.: Predictive model for the academic performance of the engineering students using CHAID and C 5.0 algorithm. Int. J. Eng. Res. Technol. **12**(6), 917–918 (2019)

Meghana, S., Trupti, P.: Consolidation of C 5.0, random forest and random trees crafts for intrusion detection system. Int. J. Res. Eng., Sci. Manage. **2**(11), 432–435 (2019)

DUBOUE, P.: The Art of Feature Engineering Essentials for Machine Learning. Cambridge University Press, Australia (2020)

Myers, R.H., Myers, S.L., Walpole, R.E.: Probability and Statistics for Engineers, 6b Ed. Prentice Hall, New York (2000)