An Analysis of What Variables Contribute to a Country's GDP

Survey

Student Name: Mondli Gina

Student ID: 0424012

COSC-4426EL Data Mining

Supervised by: Mr Sujay Kalakala

Bharti School of Engineering and Computer Science, Laurentian University

Table of Contents

Abstract

Computational economics is an interdisciplinary field that consolidates economics, computer science and mathematics. The aim is to leverage computational power to analyze economic systems and solve economic problems reliably and efficiently. Traditional statistical methods are capable of solving a limited number of problems and are prone to human error. Governments and corporations, therefore, have taken great interest in this subfield and have invested large sums of money to fund its research. This paper demonstrates the value of data analytics and machine learning algorithms in economics, with an emphasis on gross domestic product and topics in macroeconomics. The techniques used in this paper and the results they yielded were compared with those in computer science and economics literature.

Introduction

Assessing a country's economic prosperity is an enormous and complicated task. Most metrics measure a specific aspect of a country's overall performance. Gross domestic product (GDP) is a standardized economic metric that measures the total value of goods and services produced by a country within a given year (Banerjee *et al*., 2021). The most popular metric to measure human development is the Human Development Index (HDI) (Morse, 2023). Unlike GDP, HDI focuses on life expectancy, education levels and income per capita. A third metric of interest on this matter is the level of unemployment. At face value, a low unemployment level would suggest a high productivity rate and thus, a greater GDP output. Evidence shows this is often, but not always true (Cohen Kaminitz, 2023). For example, in 2022, the United States of America had an unemployment rate of 3.6% and a GDP of US$25.46 trillion. While the United Arab Emirates, enjoyed an unemployment rate of only 2.8% in 2022 (Elsayed, 2024) but had a much lower GDP of US$501.7 billion. Clearly, an analysis is required to determine what other factors contribute to a country's value output and economic growth.

Historically, GDP has been considered the go-to reference when assessing a country's performance, but researchers like Bryniuk, K. (2023) have questioned its significance and reliability. The value a country produces for the rest of the world may not necessarily reflect an increased standard of living for its citizens. I hypothesize that HDI has a strong correlation with GDP, likely as a fueling factor and thus its relationship with economic prosperity will also be examined. Population size should not be overlooked as another potential fueling factor in GDP and GDP growth. A growing population increases the potential for a proportionally growing workforce. However, population growth may conflict with HDI if a government does not ensure that its country is still able to employ its citizens (Morse, 2023). The relationship between HDI and population size will be assessed. Thereafter, I will attempt to ascertain what balance of HDI and population size figures yields the highest GDP.

Classification

Sorting data instances into different categories or classes is often a necessary task in exploratory data analysis (EDA) (Nicodemo and Satorra, 2020). When a machine learning model is built to segregate data instances into one or more classes, the process is called, "classification". Because classification places data in predefined classes, it is considered a supervised learning algorithm. The model must first be trained on a labeled dataset in order for it to discover what patterns and relationships exist between features and their corresponding class labels. Once the model has been trained, it can be used to predict what class new, unseen data points will belong to. Some common classification algorithms are decision trees, random forests, Naïve Bayes, Support Vector Machines (SVMs), logistic regression and eXtreme Gradient Boosting (XGBoost) (Tan, 2021). These are described in more detail below.

- **Decision Tree:** Classifies data points by processing them through a series of questions and making iterative classifications based on the answers they give to these questions. The data point will be ultimately classified after it responds to the last question (Tarwidi *et al*., 2023).
- **Random Forest:** An ensemble of decision trees. Each tree model is trained on a random subset of the whole dataset (Tarwidi *et al*., 2023).
- **Naïve Bayes:** A group of linear probabilistic classifiers. These classifiers use Bayes' theorem and assume that the features involved in the classification are conditionally independent (Maurya, Hussain and Singh, 2021).

  Bayes' Theorem: $$P(A|B) = {[P(A|B) * P(A)]}/{P(B)}$$

- **SVM:** A versatile machine learning model that can be used for both classification and regression problems. SVMs are particularly effective in classifying data points in datasets that have more dimensions or features than data points. SVMs are used to find the hyperplane that maximizes the margin between different classes (Tan, 2021). The architecture of an SVM is diagrammed in Figure 1.
- **Logistic Regression:** Calculates the probability of a binary outcome, such as whether or not a particular email is spam. The model achieves this by using a sigmoid function (Maurya, Hussain and Singh, 2021).

  Sigmoid Function: $$\sigma(z) = {1}/{(1 + e^2)}$$

- **XGBoost:** A scalable, distributed machine learning library. The library is an improvement on the traditional gradient-boosted decision tree library. The library provides an L1 (Lasso) and L2 (Ridge) regularizing framework and allows trees to be boosted in parallel for increased efficiency (Tarwidi *et al*., 2023). The architecture of an XGBoost is diagrammed in Figure 2.

I used logistic regression in this paper due to its simplicity and ability to perform well on low-dimensional data. XGBoost was implemented as well to compare it with logistic regression, as the literature claims that XGBoost generally creates more accurate models. XGBoost also supports the objective function mean squared error, which was a function of interest in this paper.

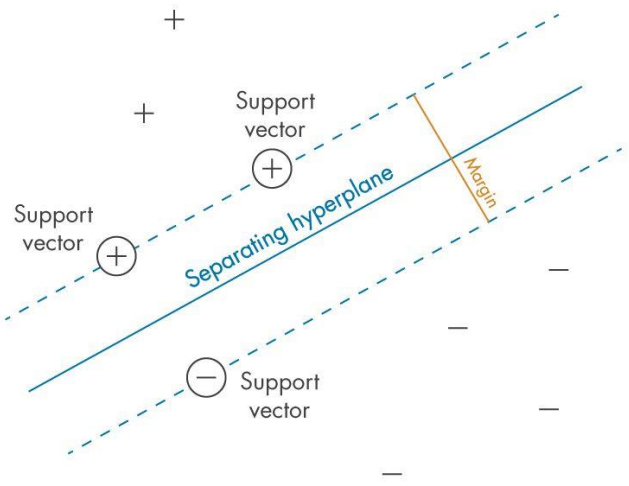

Figure 1: Defining the "margin" between classes – the criterion that SVMs seek to optimize (Paluszek and Thomas, 2017)

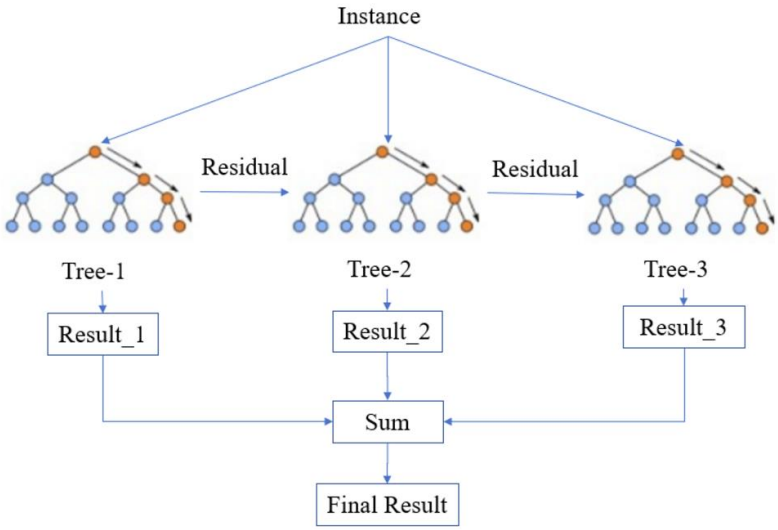

Figure 2: Simplified structure of XGBoost (Wang, Chakraborty and Chakraborty, 2020)

Correlation

Correlation is an important metric to measure the strength and direction of a relationship between two variables. Methods of correlation simply show a relationship between variables and are not enough to determine the cause of a discovered relationship (Gershman and Ullman, 2023). Some typical methods used to assess relationships between variables are Pearson, Spearman rank, point-biserial and partial correlation. The Pearson correlation is the most common method used in industry and statistics literature. This method measures the linear relationship between continuous variables on a scale of -1 to +1 (Janse *et al*., 2021). As all of the variables that were studied in this paper are continuous, this is the method that was used to measure their relationships with each other. Spearman rank is a method designed for ordinal data or data that is not normally distributed. The measurement also measures correlations on a scale of -1 to +1. Point-biserial analyzes the relationship between a continuous and a binary variable (Hashemi, Bargegol and Hamedi, 2022). None of the variables of interest were binary, hence, this method was not used. Partial correlation is similar to Pearson correlation, except it controls for the potential influence of a third variable. Although this paper did seek to observe the interaction between more than two variables and how they influence GDP together, I opted to use XGBoost for this task due to its superior computational power.

Applications of Computational Economics

XGBoost in Economic Forecasting

XGBoost, among other machine learning algorithms, has been used to make reasonable economic predictions for individuals, corporations and even governments. Macroeconomic forecasting in particular has benefitted from XGBoosting (Zuo *et al.*, 2023). Due to its ability to handle complex and intertwining relationships in large datasets, the algorithm has been used to forecast economic indicators like GDP growth, as is done in this paper, inflation rates and unemployment rates (Giraldo *et al*., 2023). Another strength of XGBoost is its ability to make predictions with limited data. Time series analyses, for example, typically require large amounts of data to ensure reliable and consistent results. XGBoost can be coupled with a time series analysis when there is limited data to help provide more reliable results. The algorithm is typically more accurate than traditional statistical models and can even provide insights into the relative importance of different economic factors. Economists may struggle to ascertain how factors like HDI and GDI contribute to a single variable of interest like GDP.

XGBoost is also less sensitive to noise and outliers in data, making it effective in real-world economic situations that may contain inaccuracies and inconsistencies in data (Tan, 2021).

SVMs in Governments

Some applications of SVMs in governmental economics include Policy evaluation, fraud detection and resource allocation (Zhang, 2022). Successful politicians usually design their policies with the guidance of an economist or data analyst. An SVM can be used to evaluate the effect of government policies on various economic conditions. After assessing the effects of these policies, SVMs can be used to optimize policy parameters and design policies that achieve desired economic goals more effectively. An example of effective SVM use in government would be tax rates. An SVM can be used to fine-tune tax rates for different demographics in order to acquire the maximal tax revenue without harming private businesses. Citizens may attempt to evade taxes or commit fraud in welfare programs. SVMs have been used to detect this kind of fraudulent activity (Amaya-Tejera *et al*., 2024). When allocating resources to education, healthcare, natural disaster aid or infrastructure, SVMs can help economists and politicians determine how to allocate funds to these different departments optimally.

Correlations between Different Condition Metrics

Pearson and Spearman's correlations are particularly interesting to economists due to their simplicity and ability to provide information about the strength and direction of different correlations (Fikri Ikhwan *et al*., 2024). Relationships that may be of interest to economists or finance professionals are interest rates and investment, inflation and unemployment or GDP growth and market performance. An investor may observe strong correlations between certain assets and opt to diversify their portfolio so as to reduce risk.

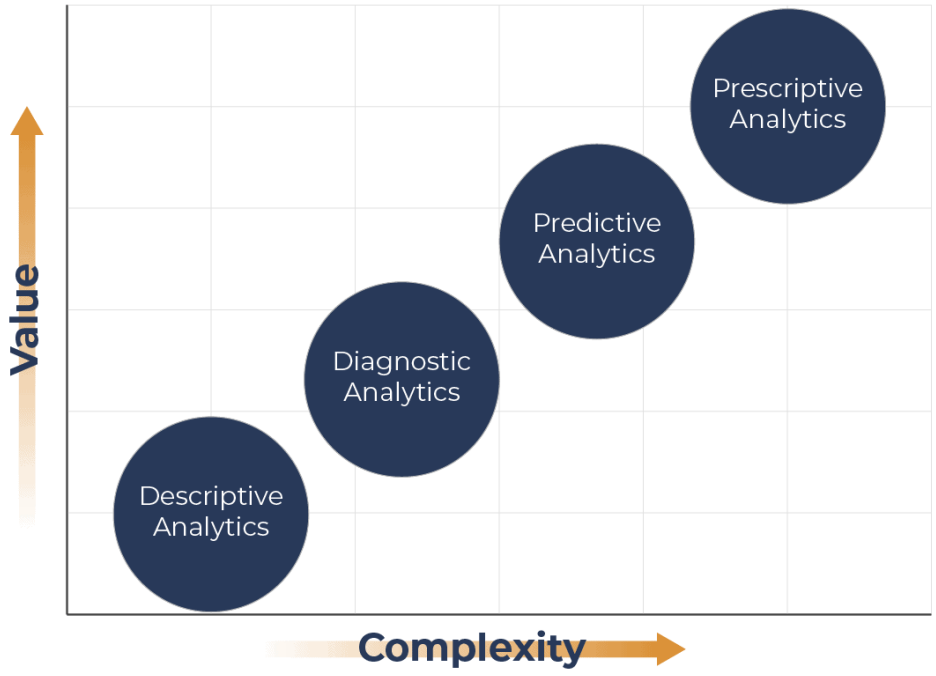

Figure 3: Evolution of data analytics in economics (Kalsbeek, 2022)

Typical Datasets Sourced in Computational Economics

Macroeconomic Datasets

As mentioned, macroeconomics features prominently in both computational economics literature and industry. Governments monitor data on national accounts to make forecasts and decisions concerning housing, public education and immigration. When conducting these studies, governments collect data on GDP, gross national product (GNP), consumption, investment and government spending data. As these are enormous datasets with potentially millions of entries, data analytics and data science techniques are necessary (Gaye, Zhang and Wulamu, 2021). Exchange rates and trade data are also of great interest to governments. Changing exchange rates between two countries may affect one country's private sector more than another. Businesses in one country may import fewer goods from another if the currency exchange rate does not favor them. It may be the case that the country with the weaker currency is simultaneously being impacted by tariffs or sanctions. Governments may use regression analyses or SVMs to identify exactly which conditions are affecting their private sector more negatively (Jun, 2021).

Microeconomic Datasets

Even though the relevance of computational economics is more obvious in macroeconomics, its applications in microeconomics should not be neglected. Medium to large corporations will be interested in firm-level and panel data (Burger *et al.*, 2024). As companies grow, their sales, profits, costs and investments will become more varied and difficult to keep track of. Complicated data like this will benefit from techniques in data analytics. Longitudinal data, such as cohort spending habits, will be of particular interest to growing companies (Gaye, Zhang and Wulamu, 2021). Lasagna plots and descriptive statistics can be used to monitor these variables over time.

Prediction Algorithms

Unsupervised Learning

Unsupervised learning algorithms or machine learning algorithms seek to recognize patterns and discover insights in unlabeled data without human supervision. Unsupervised learning models come in three main types: clustering, association rules, and dimensionality reduction. Clustering, as the name suggests, groups similar data into categories, either by K-means, hierarchical, or Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Zubair *et al.*, 2022). Some common use cases for clustering are market and image segmentation, document classification, and anomaly detection (Watson, 2023). Association rules algorithms like apriori find patterns within a set of data items in a dataset (Li *et al*., 2021). Dimensionality reduction techniques are used to reduce the number of features in a dataset. Data that is represented in a lower dimension, with only the most relevant features included, is simpler and easier to analyze. Reducing dimensionality reduces computational cost and also staves off the curse of dimensionality.

Supervised Learning

Supervised learning algorithms or machine learning algorithms use labeled data and human intervention to train machine learning models to predict results or patterns in a dataset. The algorithms can either make their predictions using regression or classification techniques. Regression-based learning models observe the relationship between dependent and independent numerical variables. Many regression algorithms exist, including but not limited to Ridge Regression, Lasso Regression, Support Vector Regression, and Gradient Boosting Machines (Jun, 2021). Once the model has been tested on a dataset, its performance can be evaluated using metrics like Mean Absolute Error and Mean Squared Error. In classification-based learning models, data points or instances are assigned to one of a number of predefined classes in a labeled dataset. Decision trees, SVMs (Amaya-Tejera *et al*., 2024), Naïve Bayes and random forests are common algorithms used to classify data points. In this report, I used an XGBoost machine to predict a country's GDP. XGBoost

was appropriate for this project because of its scalability, flexibility in handling classification and regression problems, and ability to handle enormous datasets (Lv *et al*., 2021).

Conclusion

As discussed, technology and sophisticated machine learning algorithms have propelled industrial and academic economics. SVMs, XGBoost and computational statistics offer opportunities to explore more complicated economic problems. These methods provide greater accuracy and speed than traditional statistics. The specific implementations of the algorithms mentioned in this survey are discussed in more detail in the accompanying report. Classification and supervised machine learning are more popular in computational economics literature than unsupervised machine learning because of their convenience and accuracy. Most computational economics problems can be solved more effectively with these algorithms. For this reason, particular focus is given to supervised learning and correlation measures in the report.

Bibliography

Amaya-Tejera, N., Gamarra, M., Vélez, J.I. and Zurek, E. (2024). A distance-based kernel for classification via Support Vector Machines. *Frontiers in artificial intelligence*, [online] 7. doi:https://doi.org/10.3389/frai.2024.1287875.

Burger, A., Jaklič, A., Knez, K., Kotnik, P. and Rojec, M. (2024). Firm-Level, Macroeconomic, and Institutional Determinants of Firm Growth: Evidence From Europe. *Economic and Business Review*, [online] 26(2), pp.81–103. doi:https://doi.org/10.15458/2335-4216.1336.

Fikri Ikhwan, M., Mansor, W., Khan, Z., Khairil, M., Mahmood, A., Bujang, A. and Haddadi, K. (2024). Pearson Correlation and Multiple Correlation Analyses of the Animal Fat S-Parameter. *TEM Journal*, [online] 13(1), pp.155–160. doi:https://doi.org/10.18421/TEM131-15.

Gaye, B., Zhang, D. and Wulamu, A. (2021). Improvement of Support Vector Machine Algorithm in Big Data Background. *Mathematical Problems in Engineering*, [online] 2021, p.e5594899. doi:https://doi.org/10.1155/2021/5594899.

Gershman, S.J. and Ullman, T.D. (2023). Causal implicatures from correlational statements. *PLoS One*, [online] 18(5). doi:https://doi.org/10.1371/journal.pone.0286067.

Giraldo, C., Giraldo, I., Gomez-Gonzalez, J.E. and Uribe, J.M. (2023). An explained extreme gradient boosting approach for identifying the time-varying determinants of sovereign risk. *Finance research letters*, [online] 57, pp.104273–104273. doi:https://doi.org/10.1016/j.frl.2023.104273.

Hashemi, H., Bargegol, I. and Hamedi, G.H. (2022). Using logistic regression and point-biserial correlation, an investigation of pedestrian violations and their opportunities to cross at signalized intersections. *IATSS Research*, [online] 46(3). doi:https://doi.org/10.1016/j.iatssr.2022.05.002.

Janse, R.J., Hoekstra, T., Jager, K.J., Zoccali, C., Tripepi, G., Dekker, F.W. and van Diepen, M. (2021). Conducting correlation analysis: Important limitations and pitfalls. *Clinical Kidney Journal*, [online] 14(11), pp.2332–2337. doi:https://doi.org/10.1093/ckj/sfab085.

Jun, Z. (2021). The Development and Application of Support Vector Machine. In: *Journal of Physics: Conference Series*. [online] p.052006. doi:https://doi.org/10.1088/1742-6596/1748/5/052006.

Kalsbeek, R. (2022). *Understanding the Different Types of Analytics: Descriptive, Diagnostic, Predictive, and Prescriptive*. [online] https://iterationinsights.com/. Available at: https://iterationinsights.com/article/understanding-the-different-types-of-analytics/ [Accessed 1 Dec. 2024].

Li, Z., Li, X., Tang, R. and Zhang, L. (2021). Apriori Algorithm for the Data Mining of Global Cyberspace Security Issues for Human Participatory Based on Association Rules. *Frontiers in Psychology*, [online] 11. doi:https://doi.org/10.3389/fpsyg.2020.582480.

Lv, C.-X., An, S.-Y., Qiao, B.-J. and Wu, W. (2021). Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model. *BMC Infectious Diseases*, [online] 21(1). doi:https://doi.org/10.1186/s12879-021-06503-y.

Maurya, L.S., Hussain, M.S. and Singh, S. (2021). Developing Classifiers through Machine Learning Algorithms for Student Placement Prediction Based on Academic Performance. *Applied Artificial Intelligence*, [online] 35(6), pp.403–420. doi:https://doi.org/10.1080/08839514.2021.1901032.

Nicodemo, C. and Satorra, A. (2020). Exploratory data analysis on large data sets: The example of salary variation in Spanish Social Security Data. *BRQ Business Research Quarterly*, [online] 25(3), p.234094442095733. doi:https://doi.org/10.1177/2340944420957335.

Paluszek, M. and Thomas, S. (2017). *MATLAB machine learning*. [online] New York, N.Y.] Apress. Available at: https://www.mathworks.com/campaigns/offers/next/machine-learning-with-matlab.html [Accessed 1 Dec. 2024].

Tan, H. (2021). Machine Learning Algorithm for Classification. *Journal of Physics: Conference Series*, [online] 1994(1), p.012016. doi:https://doi.org/10.1088/1742-6596/1994/1/012016.

Tarwidi, D., Pudjaprasetya, S.R., Adytia, D. and Apri, M. (2023). An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach. *MethodsX*, [online] 10, p.102119. doi:https://doi.org/10.1016/j.mex.2023.102119.

Wang, W., Chakraborty, G. and Chakraborty, B. (2020). Predicting the Risk of Chronic Kidney Disease (CKD) Using Machine Learning Algorithm. *Applied Sciences*, [online] 11(1), p.202. doi:https://doi.org/10.3390/app11010202.

Watson, D. (2023). On the Philosophy of Unsupervised Learning. *Philosophy & Technology*, [online] 36(2). doi:https://doi.org/10.1007/s13347-023-00635-6.

Zhang, Z. (2022). Prediction of Economic Operation Index Based on Support Vector Machine. *Mobile Information Systems*, [online] 2022, pp.1–11. doi:https://doi.org/10.1155/2022/3232271.

Zubair, Md., Iqbal, MD.A., Shil, A., Chowdhury, M.J.M., Moni, M.A. and Sarker, I.H. (2022). An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling. *Annals of Data Science*, [online] 11, pp.1525–1544. doi:https://doi.org/10.1007/s40745-022-00428-2.

Zuo, J., Bao, C., Meng, Q. and Zheng, Q. (2023). A Study on the Incremental Size of Social Financing Based on XGBoost and SHAP. *Procedia Computer Science*, [online] 221. doi:https://doi.org/10.1016/j.procs.2023.08.121.