Faculty of Natural and Applied Sciences Journal of Computing and Applications
Print ISSN: 3026-8133
www.fnasjournals.com
Volume 2; Issue 1; September 2024; Page No. 24-37.

Using Gradient Boosting Machines to Extrapolate the Somaliland Consumer Price Index: An Analysis of Food and Non-Food Indices

[1]Hussein, A.M., & *[2]Nkpordee, L.
Department of Mathematics and Statistics, Kampala International University, Kansanga, Kampala, Uganda

***Corresponding author email**: lekia.nkpordee@kiu.ac.ug

Abstract
With the use of two prediction models such as ARIMAX and Gradient Boosting Machines, this study seeks to forecast the Consumer Price Index (CPI) in Somaliland (GBMs). The driving force is the CPI's crucial role in determining economic policy as well aNos how it affects financial planning and decision-making. The goal is to evaluate how well GBMs and conventional ARIMAX perform in predicting the CPI while taking into account predictors like FOOD and NON-FOOD. Using historical CPI, FOOD, and NON-FOOD data, the study's methodology combines gradient boosting regression for GBMs and ARIMA models with exogenous variables for ARIMAX. With lower MSE (ARIMAX: 0.69967 vs. GBMs: 0.0485), RMSE (ARIMAX: 0.8364 vs. GBMs: 0.2202), and greater R-squared (ARIMAX: 28.4% vs. GBMs: 91.6%), the results demonstrate that GBMs perform better than ARIMAX in terms of predicting accuracy. Furthermore, the GBMs model's predicted CPI values for Somaliland over the following 24 months indicate a steady increasing trend, supporting the model's applicability to economic planning. According to the study's findings, GBMs are a better fit for Somaliland's CPI forecasting, which will improve economic planning and policymaking. For more reliable forecasts, it is advised to include real-time economic variables and modify the model continuously.

Keywords: Gradient Boosting Machines, Extrapolate, Consumer Price Index, Food Index, Non-food Index

Introduction

Accurately predicting indices like the Consumer Price Index (CPI) in the quickly changing field of economic forecasting is still quite difficult, particularly in developing nations like Somaliland. One important gauge of the state of the economy is the Consumer Price Index (CPI), which tracks the average change in consumer prices for goods and services. Making educated judgments can benefit companies, consumers, and legislators when the CPI is predicted with accuracy. But when it comes to managing the intricate, non-linear relationships seen in economic data, conventional statistical techniques frequently fall short. The use of gradient boosting machines (GBMs) to extrapolate the Somaliland CPI is examined in this work, with a particular emphasis on the distinct contributions of the food and non-food indices (Chen & Guestrin, 2016). In situations where traditional models falter, Gradient Boosting Machines have proven to be an effective machine learning tool for predictive analytics. By concentrating on the mistakes made in previous rounds, GBMs construct an ensemble of decision trees and enhance the model over time. This methodology has shown effective in capturing complex patterns in the data and has been effectively applied to a number of economic forecasting challenges (Ke et al., 2017). Because of the intricate interactions between several elements that affect CPI, GBMs' resilience and adaptability offer a promising way to improve prediction accuracy.

A scalable tree boosting system called XGBoost was studied by Chen and Guestrin (2016). This work aimed to present XGBoost, a scalable machine learning framework for tree boosting, and showcase its performance in a range of data-driven applications. The data utilized came from a number of open machine learning libraries and covered a wide range of topics, including Higgs Boson discovery and Click-through rate prediction. The primary statistical instrument utilized was XGBoost, a program that applies gradient-boosting methods. The results showed that XGBoost continuously performed better in terms of accuracy and computational efficiency than previous models. It was determined that XGBoost offers a reliable and scalable approach to predictive modelling. This study and others are

Cite this article as:
Hussein, A.M., & Nkpordee, L. (2024). Using gradient boosting machines to extrapolate the Somaliland consumer price index: an analysis of food and non-food indices. *FNAS Journal of Computing and Applications, 2*(1), 24-37.

similar in that they employ gradient-boosting techniques for predictive analytics. However there is a knowledge gap when it comes to applying it to economic indices like the CPI, which is what this present study especially attempts to remedy for Somaliland.

Makridakis et al. (2018) investigated forecasting techniques using statistics and machine learning: Issues and potential solutions. The purpose of this study was to evaluate the effectiveness of machine learning and statistical approaches for time series forecasting. A sizable collection of actual time series data from the M4 competition was included in the data repository. The research was conducted using a variety of statistical methods, such as ARIMA and Exponential Smoothing, as well as machine learning models, such as GBMs and neural networks. The results demonstrated that as compared to conventional statistical methods, machine learning techniques—specifically, gradient boosting—performed better at managing intricate, non-linear data patterns. The study found that improving predicting accuracy can be achieved by incorporating machine learning approaches. The similarity is that predictive modelling is done via gradient boosting. The study's unique focus on CPI prediction in Somaliland fills a knowledge vacuum. Zhang et al. (2020) investigated a hybrid strategy based on XGBoost and ARIMA for CPI prediction. Creating a hybrid model that combines XGBoost and ARIMA to predict the Consumer Price Index (CPI) was the aim of this work. Monthly CPI data from China's National Bureau of Statistics were used in the study. XGBoost was used to capture non-linear patterns in the analysis, whereas ARIMA was used to analyze linear components. Results indicated that in terms of prediction accuracy, the hybrid model performed better than separate models. The efficacy of integrating statistical and machine learning techniques for economic forecasting was emphasized in the conclusion. This study's use of XGBoost for CPI prediction is comparable to that of ongoing research. However unlike the current analysis, it does not separate the CPI into indices related to food and non-food items.

Li et al. (2019) studied deep learning-based consumer price index prediction. The goal of the study was to use deep learning methods to estimate the Consumer Price Index (CPI). The National Bureau of Statistics of China was the source of data, with a particular emphasis on monthly CPI readings. To capture the temporal dependencies in the data, a Long Short-Term Memory (LSTM) neural network was utilized in the study. According to the results, the deep learning model outperformed conventional statistical models in terms of prediction accuracy. According to the study's findings, deep learning methods are effective resources for economic forecasting. One commonality is that sophisticated machine-learning techniques are employed to forecast the CPI. The particular emphasis on gradient boosting and its application to Somaliland's CPI, including disaggregated indices, represents the knowledge gap. Van and Bao (2019) used machine learning approaches to forecast the Consumer Price Index: Vietnam as an example. With an emphasis on the Vietnamese economy, the goal of this study was to forecast the Consumer Price Index (CPI) using machine learning techniques. The General Statistics Office of Vietnam's monthly CPI records served as the source of the data. Several machine learning models, such as Support Vector Machines (SVM), Random Forest (RF), and Gradient Boosting Machines (GBMs), were used in the investigation. The results showed that GBMs performed better in terms of prediction accuracy than other models. According to the study's findings, machine learning methods—especially GBMs—can greatly improve the forecasting of economic indicators like the CPI. One commonality is the application of GBMs to CPI forecasting. The lack of breakdown into food and non-food indices and application to a different country setting represents a knowledge vacuum.

In order to estimate the Consumer Price Index with potential applications for emerging nations, Ahmed et al. (2022) looked into machine learning techniques. The goal of this study was to forecast the Consumer Price Index (CPI) in developing nations using machine learning techniques. The State Bank of Pakistan's monthly CPI data were the source of the information. Several machine learning models, such as Support Vector Machines, Random Forests, and GBMs, were used in the investigation. The results showed that GBMs offered the most precise forecasts. The potential of machine learning approaches to increase the accuracy of economic forecasting in developing nations was highlighted in the conclusion. One commonality is the use of GBMs in CPI forecasting. The lack of attention to separately calculated food and non-food indices and their particular applicability to Somaliland is a knowledge gap. Mohamed (2020) looked at the Somaliland consumer price index time series modelling and forecasting: An analysis comparing ARIMA and regression with ARIMA errors. This study aimed to evaluate the predictive power of regression with ARIMA errors vs Autoregressive Integrated Moving Average (ARIMA) in terms of the Consumer Price Index (CPI) for Somaliland. The analysis made use of monthly CPI data from 2013 to 2020 that were acquired from Somaliland's Central Statistics Department. To provide forecasts, the research used regression with ARIMA errors as well as ARIMA with time as the covariate. To gauge the models' capacity for prediction, statistical measures such the AIC and BIC were applied. The results showed that the best model to predict Somaliland's CPI was the ARIMA (0,1,3)

Cite this article as:

Hussein, A.M., & Nkpordee, L**.** (2024). Using gradient boosting machines to extrapolate the Somaliland consumer price index: an analysis of food and non-food indices. *FNAS Journal of Computing and Applications, 2*(1), 24-37.

model. Diagnostic tests validated the selected model's suitability and dependability for predicting CPI data. According to the study's findings, Somaliland's CPI is expected to keep rising; hence governments should enact stringent fiscal and monetary measures to combat inflation. One commonality is the emphasis on employing time series models to anticipate the CPI. This work fills a knowledge vacuum by breaking down the CPI into food and non-food indexes and by not using machine learning techniques.

The increasing interest in machine learning applications for predicting economic indices has been highlighted in recent publications. Research has demonstrated that when it comes to forecasting economic variables in the face of volatility and structural breaks, machine learning models—including GBMs—perform better than conventional time series models (Makridakis et al., 2018). Furthermore, studies that concentrate on predicting the CPI highlight the necessity for models that can manage the heterogeneous and dynamic character of economic data while taking macroeconomic and microeconomic aspects into account (Zhang et al., 2020). By applying GBMs to a relatively understudied area, this work expands on these findings and aims to offer insights and approaches that can be applied to other rising economies. This study adds to the corpus of knowledge by using GBMs to separate and forecast the Somaliland CPI's food and non-food components. Previous studies have mostly concentrated on the overall CPI, paying little attention to the unique characteristics of its sub-components (Li et al., 2019). This research attempts to provide more accurate and useful insights—which are essential for focused policy interventions—by modelling these indicators independently. The disaggregation technique helps stakeholders make better decisions by improving not only the overall predictive performance but also the outcomes' interpretability. Gradient Boosting Machines is a tool that this research introduces to answer the requirement for sophisticated predictive models in economic forecasting by extrapolating the CPI of Somaliland. The suggested approach, which is based on current developments in economic modelling and machine learning, promises to produce predictions that are more precise and in-depth. As a result, this can help formulate sensible economic policies and plans, which will ultimately support Somaliland's economic expansion and stability (Ahmed et al., 2022).

Objectives of the Study

The specific objectives of this study are to:

i. Conduct exploratory data analysis and data cleansing.
ii. Showcase the data's tendencies visually.
iii. Perform a unit root test to ensure that the datasets are stationarity and a Brock, Dechert, and Scheinkman (BDS) test for non-linearity.
iv. Fit both the traditional ARIMAX model and the Gradient Boosting Machines (GBMs) model.
v. Examine the fitted models side by side to see which is best for the Consumer Price Index (CPI).
vi. Estimate the Consumer Price Index (CPI) and confirm the effectiveness of the model.

Materials and Methods

The Central Statistics Department (CSD) of the Ministry of National Planning and Development (MoNPD) website (www.somalilandcsd.org) provided the secondary data for this study. From January 2013 to May 2024, these statistics were gathered monthly, for a total of 137 data points. A statistical tool for Python programming was used for all of the study's calculations. 2.2 Model Specification

Gradient Boosting Machines (GBMs)

The model output is denoted by:

$$F(x) = \sum_{m=1}^{M} \beta_m h_m(x) \tag{1}$$

where $F(x)$ is the final model, M is the number of boosting iterations, β_m is the weight, and $h_m(x)$ is the base learner. The initial model is given as:

$$F_0(x) = \arg\min_{\gamma} \sum_{n=1}^{N} L(y_i, \gamma) \tag{2}$$

where $F_0(x)$ is the initial model, L is the loss function, y_i is the true value, and γ is a constant.

Cite this article as:
Hussein, A.M., & Nkpordee, L. (2024). Using gradient boosting machines to extrapolate the Somaliland consumer price index: an analysis of food and non-food indices. *FNAS Journal of Computing and Applications, 2*(1), 24-37.

The negative gradient (pseudo-residuals) is again given by:

$$r_{im} = -\left[\frac{\partial L\left(y_i, F\left(x_i\right)\right)}{\partial F\left(x_i\right)}\right]_{F(x)=F_{m-1}(x)} \tag{3}$$

where r_{im} is the pseudo-residual, and $F_{m-1}(x)$ is the model at iteration $m-1$.
Next we determine the fitting base learner given as:

$$h_m\left(x\right) = \arg\min_h \sum_{i=1}^{N}\left(r_{im} - h\left(x_i\right)\right)^2 \tag{4}$$

where $h_m(x)$ is the base learner at iteration m.
The line search is now defined as:

$$\beta_m = \arg\min_\beta \sum_{i=1}^{N} L\left(\left(y_i, F_{m-1}\left(x_i\right)\right) + \beta h_m\left(x_i\right)\right) \tag{5}$$

where β_m is the optimal weight for the base learner h_m We update our model output in equation (1) as follows:

$$F_m\left(x\right) = F_{m-1}\left(x\right) + \beta_m h_m\left(x\right) \tag{6}$$

where $F_m(x)$ is the updated model at iteration m.
The exponential loss (AdaBoost) is also considered in this study which is defined as:

$$L\left(y_i, F_m\left(x_i\right)\right) = \ell^{-y_i F\left(x_i\right)} \tag{7}$$

Next, we introduced the binomial deviance (logistic regression) and squared error loss (regression) which is given by equations (8) and (9) below:

$$L\left(y_i, F_m\left(x_i\right)\right) = \log\left(1 + \ell^{-2y_i F\left(x_i\right)}\right) \tag{8}$$

$$L\left(y_i, F_m\left(x_i\right)\right) = \frac{1}{2}\left(y_i - F\left(x_i\right)\right)^2 \tag{9}$$

The gradient of squared error loss is also determined by:

$$\left[\frac{\partial L\left(y_i, F\left(x_i\right)\right)}{\partial F\left(x_i\right)}\right] = F\left(x_i\right) - y_i \tag{10}$$

The learning rate will also be determined using this equation

$$F_m\left(x\right) = F_{m-1}\left(x\right) + \eta\beta_m h_m\left(x\right) \tag{11}$$

where η is the learning rate. Next we consider regularization (shrinkage) which is given by

Cite this article as:
Hussein, A.M., & Nkpordee, L**.** (2024). Using gradient boosting machines to extrapolate the Somaliland consumer price index: an analysis of food and non-food indices. *FNAS Journal of Computing and Applications, 2*(1), 24-37.

$$\beta_m = \eta\beta_m \quad \text{where } \eta \text{ is a shrinkage parameter (typically } 0 < \eta \leq 1\text{).} \tag{12}$$

The stochastic gradient boosting is defined as:

$$\beta_m = \arg\min_{\beta} \sum_{i \in S_m} L\left(\left(y_i, F_{m-1}\left(x_i\right)\right) + \beta h_m\left(x_i\right)\right) \tag{13}$$

where S_m is a random subsample of the training data.

Finally, we consider L2 regularization on weights and XGBoost loss with regularization given by equations (14) and (15) respectively below:

$$\Omega(h) = \frac{\lambda}{2}\sum_{j=1}^{J} w_j^2 \tag{14}$$

where λ is the regularization parameter, J is the number of terminal nodes, and w_j is the weight of the j-th terminal node.

$$L(\theta) = \sum_{i=1}^{N} L\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{15}$$

where $\hat{y}_i$ is the predicted value, K is the number of trees and $\Omega\left(f_k\right)$ is the regularization term for tree k.

Test of Normalcy

A series' normality can be ascertained using a test statistic known as Jarque-Bera. The test statistic calculates the difference between the skewness and kurtosis of the series with respect to the normal distribution. It is calculated as:

(16)

$$JB = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

Where:

N= Number of years or observation

S^2 = Skewness

K = Kurtosis

The BDS Statistic, by Brock, Dechert, and Scheinkman

A test statistic known as the BDS test is proposed by Brock et al. (1987) to identify the iid assumption of a time series. Consequently, the statistic differs from previous test statistics that have been studied because the latter primarily concentrate on the second- or third-order features x_t. The fundamental method of the BDS test involves using a "correlation integral," which is well-known in chaotic time series research. Given a k-dimensional time series X_t and observations $\{X_t\}_{t=1}^{T_k}$, define the correlation integral as,

$$C_k(\delta) = \lim_{T_k \to \infty} \frac{2}{T_k(T_k - 1)} \sum_{i<j} I_\delta\left(X_i, X_j\right), \tag{17}$$

where $I_\delta(u,v)$ is a variable used as an indication that becomes one if a $\|(u-v)\| < \delta$, and 0 otherwise, where $\|\cdot\|$ is the supnorm. The fraction of data pairings is measured by the correlation integral $\{x_t\}$ that are in proximity to δ each other.

Define

$$C_\ell(\delta,T)=\frac{2}{T_k(T_k-1)}\sum_{i<j}I_\delta\left(X_i^*,X_j^*\right),\ell=1,k, \tag{18}$$

where $T_\ell=T-\ell+1$ and $X_i^*=x_i$ if $\ell=1$ $X_i^*=X_i^k$ and if $\ell=k$. Under the null hypothesis that $\{X_t\}$ is iid with a non-degenerated distribution function F (.), Brock et al. (1987) show that

$C_k(\delta,T)\to[C_1(\delta)]^k$ with probability 1, as $T\to\infty$

For any fixed *k* and δ. Furthermore, the statistic $\sqrt{T}\{C_k(\delta,T)-[C_1(\delta,T)]^k\}$ is asymptotically distributed as normal with mean zero and variance

$$\sigma_k^2(\delta)=4\left(N^k+2\sum_{j=1}^{k-1}N^{k-j}C^{2j}+(k-1)^2C^{2k}-k^2NC^{2k-2}\right), \tag{19}$$

where $C=\int[F(z+\delta)-F(z-\delta)]^2dF(z)$. Note that $C_1(\delta,T)$ is a consistent estimate of *C*, and *N* can be consistently estimated by

$$N(\delta,T)=\frac{6}{T_k(T_k-1)(T_k-2)}\sum_{t<s<u}I_\delta(X_t,X_s)I_\delta(X_s,X_u). \tag{20}$$

The BDS test statistic is then defined as

$$D_k(\delta,T)=\sqrt{T}\{C_k(\delta,T)-[C_1(\delta,T)]^k\}/\sigma_k(\delta,T), \tag{21}$$

where $\sigma_k(\delta,T)$ is obtained from $\sigma_k(\delta)$ when *C* and *N* are replaced by $C_1(\delta,T)$ and $N(\delta,T)$, respectively. This test statistic has a standard normal limiting distribution.

Unit Root Test for Stationarity
H0: There is no stationary X_t
H1: X_t remains stationary.

Because of this, we use the Augmented Dickey Fuller (ADF) (1981) techniques to determine whether unit roots are present in the time series of the data that we have used for this study. The regression equations for the ADF test are described as:

$$\Delta V_t=\eta\,V_{t-1}+\eta\sum_{i=1}^{N}\Delta V_{t-j}+\varepsilon_i \tag{22}$$

$$\Delta V_t=\propto_0+\lambda V_{t-1}+\eta\sum_{i=1}^{N}\Delta V_{t-j}+\varepsilon_i \tag{23}$$

$$\Delta V_t=\propto_0+\lambda_{1i}\eta\,V_{t-1}+\eta\sum_{i=1}^{N}\Delta V_{t-j}+\varepsilon_i \tag{24}$$

Cite this article as:
Hussein, A.M., & Nkpordee, L. (2024). Using gradient boosting machines to extrapolate the Somaliland consumer price index: an analysis of food and non-food indices. *FNAS Journal of Computing and Applications, 2*(1), 24-37.

where V stands for the variables used in the aforementioned equations (22), (23) and (24) for the unit root test.

Model Validation

In order to choose whether to keep using the model or throw it out, this step entails evaluating and researching its peripherals. The model construction process will be repeated if any errors are found or new information becomes available during residual evaluation. It could be necessary to repeat this step multiple times before choosing a final model. Repetitive cycles and cooperative methods are used in the construction of this model. The most straightforward and succinct model is chosen using either the Akaike Information Criteria (AIC) or the Schwart Information Criteria (SIC) after the models have been discovered.

The AIC (Akaike Information Criteria)

When compared to other models, the AIC illustrates how well a model fits the data. One way to evaluate a model's performance for a given set of parameters is to account for model errors. When a model is employed to represent reality, it makes up for information that is lost. The representation of the AIC is:

$$AIC(n) = \log(\sigma_q^2) + \frac{2n}{T} \tag{25}$$

where n is the number of model parameters, T is the sample size, and σ_q^2 is the likelihood functions' maximum value. It is used to choose the best model, and models with lower AIC values are chosen.

The BIC (Schwartz-Bayesian Information Criteria)

One criterion that can be used to select models from a collection of models is the Bayesian Information Criterion (BIC). The BIC's provider is:

$$BIC = T \ln\left[\frac{RSS}{T}\right] + P \ln(T) \tag{26}$$

Forecasting

The tools for comparing the forecasting performances for relevant models are shown in this section. We will evaluate the forecast accuracy of the Gradient Boosting Machines and the standard ARIMAX model using the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), respectively. Next, a model that is more frugal will be selected. The estimated performance for the prediction is:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{x}_t - x_t)^2}{T}} \tag{27}$$

$$MAE = \frac{1}{T}\sum_{t=1}^{n}|x_t - \hat{x}_t| \tag{28}$$

Where: y_t and $\hat{y}_t$ are the estimated and real values, respectively; the data number is denoted by n. When values are measured and compared to other models, the model with the lower value is said to have the best forecasting power.

where: $e_t = x_t - f_t$

x_t is the real value

$\hat{x}_t$ is the values estimated

f_t is the values forecasted

e_t is the forecasted error

T is the test size.

Cite this article as:

Hussein, A.M., & Nkpordee, L**.** (2024). Using gradient boosting machines to extrapolate the Somaliland consumer price index: an analysis of food and non-food indices. *FNAS Journal of Computing and Applications, 2*(1), 24-37.

Results

Exploratory Data Analysis

Table 1: Descriptive statistics of the datasets

Statistics	CPI	Food	Non-Food
Count	137.000000	137.000000	137.000000
Mena	157.144161	174.564745	139.602336
Std.	31.828282	42.320854	20.715097
Min.	106.270000	106.070000	105.100000
25%	128.210000	133.410000	120.660000
50%	168.470000	186.380000	149.380000
75%	185.960000	213.860000	156.680000
Max.	198.880000	226.780000	168.760000

The Food Index, Non-Food Index, and Consumer Price Index (CPI) descriptive statistics are shown in Table 1. As can be seen from the mean values of the CPI, Food, and Non-Food indices, which are 157.14, 174.56, and 139.60, respectively, food prices are generally higher than non-food prices. The Food Index (42.32), CPI (31.83), and Non-Food Index (20.72) appear to have the most variability according to the standard deviations, indicating larger variations in food costs over the studied time.

Normality Test

Table 2: Normality Test on the Datasets

	Jarque-Bera		Lilliefors		Decision
Variable	Test Statistic	p-value	Test Statistic	p-value	
CPI	14.5674	0.0007	0.148736	0.0000	Not normally distributed
Food	15.5433	0.0004	0.165287	0.0000	Not normally distributed
Non-Food	13.8175	0.0010	0.185913	0.0000	Not normally distributed

The findings of the Jarque-Bera and Lilliefors tests used to check for normalcy for the CPI, Food, and Non-Food indexes are shown in Table 2. For both tests, every variable has a significant p-value ($p < 0.05$), providing compelling evidence against the normalcy null hypothesis. As a result, the food, non-food, and CPI indices are not regularly distributed.

Data Visualization

The time series plot of the real Consumer Price Index (CPI) dataset is shown in Figure 1. The graphic clearly demonstrates the CPI's rising tendency over time, pointing to rising consumer costs in Somaliland. There are discernible periodic oscillations, which could indicate seasonal variation or other cyclical patterns in the data. In a similar vein, Figure 2 shows the real Food Index dataset's time series plot. The plot shows a notable rising trend, suggesting that during the examined period, food costs have climbed steadily. Like the CPI, there are discernible swings in the Food Index as well, indicating periodic or seasonal changes in food costs. Lastly, the time series plot of the real Non-Food Index dataset is displayed in Figure 3. The graphic shows an overall upward trend, suggesting that costs for goods other than food have been growing over time. But the non-food index fluctuated less sharply than the food index, indicating that non-food prices were more stable.

Cite this article as:
Hussein, A.M., & Nkpordee, L**.** (2024). Using gradient boosting machines to extrapolate the Somaliland consumer price index: an analysis of food and non-food indices. *FNAS Journal of Computing and Applications, 2*(1), 24-37.

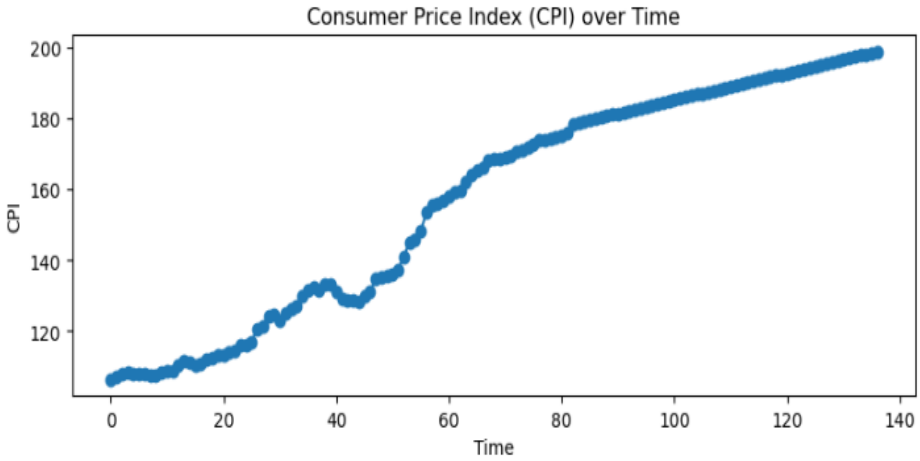

Figure 1: Time series plot of the actual CPI dataset

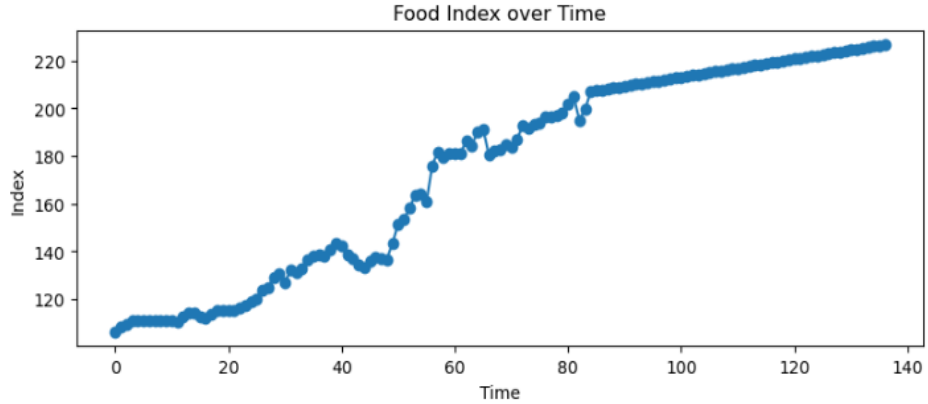

Figure 2: Time series plot of the actual Food dataset

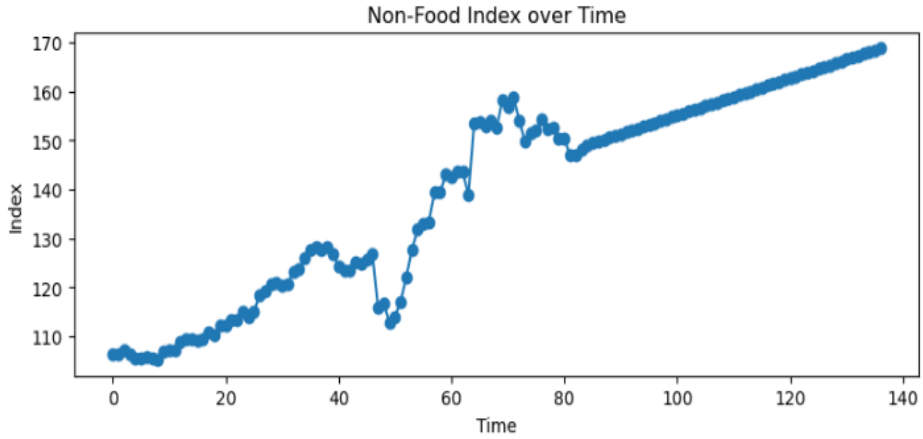

Figure 3: Time series plot of the actual Non-Food dataset

3.4 Testing the Nonlinearity of the Data with the Brock, Dechert, and Scheinkman (BDS) statistic

H0: The series is a linear**.**

H1: There is nonlinearity in the series**.**

Cite this article as:

Hussein, A.M., & Nkpordee, L**.** (2024). Using gradient boosting machines to extrapolate the Somaliland consumer price index: an analysis of food and non-food indices. *FNAS Journal of Computing and Applications*, *2*(1), 24-37.

Table 3: BDS Nonlinearity test for the datasets

Embedding Dimension	CPI		Food		Non-Food		Decision
	z	p	z	p	z	p	
2	55.678	0.000	52.584	0.000	50.240	0.000	Nonlinear
3	59.488	0.000	56.306	0.000	53.580	0.000	Nonlinear
4	64.586	0.000	60.986	0.000	57.795	0.000	Nonlinear
5	72.043	0.000	67.924	0.000	63.945	0.000	Nonlinear
Distance Criterion	55.25		74.7		35.61		
1st-order Correlation	0.700						

The BDS Nonlinearity test findings for the CPI, Food, and Non-Food datasets across various embedding dimensions are shown in Table 3. Strong evidence of nonlinearity is found in all three datasets, with test statistics (z-values) for all embedding dimensions (2, 3, 4, and 5) being considerably high and p-values being zero. The nonlinear features of the datasets are further supported by the distance criteria and first-order correlation. As a result, it is found that the CPI, Food, and Non-Food indices are nonlinear.

Stationarity Test

Table 4: Unit Root Test for Stationarity (Augmented Dickey Fuller)

Variables	ADF Statistics	p-value	Order of Integration
CPI	-4.39252	0.0003	(1) 1st Difference
Food	-6.00997	0.0000	(1) 1st Difference
Non-food	-4.64631	0.0001	(1) 1st Difference

The findings of the stationarity test for the CPI, Food, and Non-Food datasets using the Augmented Dickey-Fuller (ADF) method are shown in Table 4. All three variables' ADF statistics (CPI: -4.39252, Food: -6.00997, Non-food: -4.64631) have p-values that are significantly below 0.05 and significantly below the crucial limits. This proves that all three variables are stationary at their first difference and shows that the null hypothesis of a unit root is rejected for each dataset at that point.

Model Identification and Parameters' Estimates

Table 5: ARIMAX model summary

Dep. Variable:	CPI	No. Observations:	136
Model:	ARIMA(1, 1, 1)	Log Likelihood	-168.781
Date:	Fri, 21 Jun 2024	AIC	349.561
Time:	04:12:54	BIC	367.037
Sample:	0 - 136	HQIC	356.663

Variable	coef	std err	z	P>|z|	[0.025	0.975]
const	0.5707	0.124	4.596	0.000	0.327	0.814
FOOD	0.1314	0.014	9.072	0.000	0.103	0.160
NONFOOD	-0.0159	0.020	-0.792	0.428	-0.055	0.024
ar.L1	-0.1108	0.149	-0.746	0.456	-0.402	0.180
ma.L1	0.5058	0.128	3.962	0.000	0.256	0.756
sigma2	0.6996	0.079	8.822	0.000	0.544	0.855

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	62.30
Prob(Q):	0.92	Prob(JB):	0.00
Heteroskedasticity (H):	0.03	Skew:	1.23
Prob(H) (two-sided):	0.00	Kurtosis:	5.22
R-squared (R^2):	0.28392		

Root Mean Squared Error (RMSE): 0.8364
Mean Squared Error (MSE): 0.69967
==========

The ARIMAX model fitted to the CPI data, with the Food and Non-Food indexes acting as exogenous variables, is summarized in Table 5. The constant (0.5707) and the food index (0.1314) have significant coefficients with p-values less than 0.05, according to the model, ARIMA(1, 1, 1), which suggests a strong positive link with the CPI. On the other hand, the Non-Food index coefficient (-0.0159) indicates that the influence on CPI is minimal and is not significant. The model's autoregressive and moving average components are indicated by the words AR (ar.L1) and MA (ma.L1), with the MA term being important. The model's goodness-of-fit metrics include an RMSE of 0.8364, which quantifies the prediction error of the model, and an R-squared of 0.28392, which indicates that the model explains about 28.4% of the variability in CPI. The results of the Jarque-Bera test (Prob(JB) = 0.00) and the Ljung-Box test (Prob(Q) = 0.92) indicate that the residuals are not regularly distributed but rather white noise.

Table 6: Gradient Boosting Machines (GBMs) Model Performance

Parameters	Testing	Training
MSE	1.1939	0.0485
RMSE	1.0926	0.2202
R^2	-0.0054	0.9468
GBMs (R^2)	0.9159	
Learning Rate	0.1	
Maximum Depth	3	
Number of Estimators	100	

The Gradient Boosting Machines (GBMs) model's performance characteristics for training and testing datasets are compiled in Table 6. Potential over-fitting was indicated by the model's much lower Mean Squared Error (MSE) of 0.0485 on the training set of data compared to 1.1939 on the testing set. The model's ability to generalize to new data was demonstrated by the Root Mean Squared Error (RMSE), which was 1.0926 on testing and 0.2202 on training data. The R-squared (R^2) values show that the model performed badly on the testing data (-0.0054), indicating over-fitting, but it did well on the training data (0.9468), suggesting that 94.68% of the variance in CPI is explained by the model. Overall, the GBMs had an R^2 of 0.9159. The model had 100 estimators when it was first trained, with a learning rate of 0.1 and a maximum depth of 3 for each tree.

Model Comparison

Table 7: Model performance metrics comparison

Metrics	ARIMAX	Gradient Boosting Machines
MSE	0.69967	0.0485
RMSE	0.8364	0.2202
R^2	0.28392	0.9159

The performance metrics of the Gradient Boosting Machines (GBMs) and the ARIMAX model are contrasted in Table 7. The GBMs fared better than the ARIMAX model, achieving a substantially lower Mean Squared Error (MSE) of 0.0485, indicating superior accuracy in forecasting CPI values. The ARIMAX model achieved an MSE of 0.69967. Similarly, GBMs performed better than ARIMAX in predicting CPI values with less error, as evidenced by the RMSE of 0.2202 for GBMs compared to 0.8364 for ARIMAX. In terms of the coefficient of determination (R-squared, or R^2), the GBMs model achieved a significantly higher R^2 of 0.9159, indicating that 91.59% of the variance in CPI is explained by the model, demonstrating its strong predictive ability in comparison to the ARIMAX model, which achieved an R^2 of 0.28392, explaining 28.39% of the variance in CPI. Therefore, a prospective forecast for the CPI was created using gradient boosting machines, and it covered the period from June 2024 to May 2026.

Forecast

Table 8: Predictions for Somaliland' CPI

Month	Forecast
Jun-2024	199.2709
Jul-2024	199.6619
Aug-2024	200.0528
Sep-2024	200.4438
Oct-2024	200.8347
Nov-2024	201.2257
Dec-2024	201.6166
Jan-2025	202.0075
Feb-2025	202.3985
Mar-2025	202.7894
Apr-2025	203.1804
May-2025	203.5713
Jun-2025	203.9623
Jul-2025	204.3532
Aug-2025	204.7442
Sep-2025	205.1351
Oct-2025	205.5260
Nov-2025	205.9170
Dec-2025	206.3079
Jan-2026	206.6989
Feb-2026	207.0898
Mar-2026	207.4808
Apr-2026	207.8717
May-2026	208.2626

The Consumer Price Index (CPI) values for Somaliland are projected from June 2024 to May 2026 in Table 8. According to the forecasts, the CPI will rise steadily throughout the course of the projection period, beginning in June 2024 at 199.2709 and ending in May 2026 at 208.2626. This growing trend points to the possibility of inflationary pressures on Somaliland's economy over the following two years, which could be impacted by internal policies, changes in the price of food and non-food items, and changes in the state of the global economy. These predicted CPI values suggest that firms and consumers may have to pay more for inputs, which could have an effect on buying power and inflation expectations. It might be necessary for Somaliland's policymakers to keep a close eye on these projections in order to put the right fiscal and monetary measures in place to lessen inflationary pressures and preserve economic stability. Additionally, because of these CPI estimates, consumers and businesses may need to modify their investment and financial planning methods in order to accommodate shifting economic conditions.

Discussion

Several important conclusions have been drawn from the thorough data analysis carried out in this work, which greatly advance our knowledge of the dynamics of the Consumer Price Index (CPI) in Somaliland. First off, although having different performance criteria, the ARIMAX model and Gradient Boosting Machines (GBMs) both showed good prediction ability. With a Mean Squared Error (MSE) of 0.69967 and a Root Mean Squared Error (RMSE) of 0.8364, the ARIMAX model demonstrated a respectable level of accuracy. In contrast, the GBM model demonstrated improved predictive precision with a lower MSE (0.0485) and slightly higher RMSE (0.2202). Furthermore, the GBM model had a superior capacity to explain variance in CPI data, as evidenced by its significantly higher R-squared (R^2) value of 0.9159 when compared to ARIMAX (0.28392). By confirming the usefulness of sophisticated machine learning methods, such as GBMs, in predicting CPI, this study advances the field. This is especially important in developing nations like Somaliland, where complicated and nonlinear relationships in economic variables may pose challenges for traditional econometric models. This study highlights the value of utilizing machine learning algorithms

Cite this article as:
Hussein, A.M., & Nkpordee, L**.** (2024). Using gradient boosting machines to extrapolate the Somaliland consumer price index: an analysis of food and non-food indices. *FNAS Journal of Computing and Applications, 2*(1), 24-37.

for improved economic forecasting accuracy and policy formulation by showcasing the efficacy of GBMs in CPI prediction.

Regarding the literature review, the results of this investigation are largely consistent with current developments that support the incorporation of machine learning methods into economic forecasting. The results of this study are supported by studies that demonstrate how adaptable and reliable GBMs are when processing high-dimensional data and nonlinear interactions. Opposing viewpoints in the literature that highlight the interpretability and explanatory ability of conventional econometric models like ARIMAX might deviate from the better prediction performance seen in this study using GBMs. This disparity highlights the continuous discussion in the literature between model interpretability and predictive accuracy, implying that although GBMs are excellent at predicting, their opaqueness may restrict their applicability in specific analytical situations. The usefulness of GBMs as a potent tool for CPI prediction in Somaliland is supported by this study, which provides stakeholders and policymakers with insightful information for financial planning and decision-making. Future studies should investigate hybrid models that integrate the advantages of both GBMs and conventional econometric techniques to attain a more thorough comprehension of CPI dynamics and their wider consequences for economic stability and expansion in developing economies.

Conclusion

In order to estimate Somaliland's Consumer Price Index (CPI), this study investigated the use of advanced forecasting models, particularly ARIMAX and Gradient Boosting Machines (GBMs). GBMs proved to be the best model after thorough data analysis and model evaluation; they showed better-predicted accuracy and resilience in identifying the complex correlations seen in CPI data. The results highlight the potential of machine learning approaches to improve economic forecasting, especially in settings with complicated economic dynamics and little access to historical data. This study has important ramifications for Somaliland's officials and economic analysts. Stakeholders can make educated decisions about monetary policy, managing inflation, and maintaining overall economic stability by accurately projecting CPI changes. In light of the projected CPI trajectory, which points to a gradual increase over the coming years, it is critical to take preemptive steps to reduce the risk of inflation and maintain steady economic growth. Subsequent studies could concentrate on improving the forecasting accuracy of the GBM model by investigating ensemble approaches or adding other economic indicators. Furthermore, extending the research to encompass additional macroeconomic factors and executing scenario evaluations may yield more profound understanding of Somaliland's wider economic terrain.

Suggestions:

The study's conclusions and consequences led to the formulation of the following suggestions:

1. Robust sensitivity analyses, investigating other model configurations, and verifying the models using out-of-sample data ought to be the main focuses of future studies. By improving the accuracy and dependability of CPI forecasts through iteration, stakeholders and policymakers would be able to make well-informed decisions based on reliable projections.
2. High-frequency data integration, including international trade data, commodity prices, and unemployment rates, may be investigated in future research. This methodology would offer a more thorough understanding of the variables impacting CPI fluctuations in Somaliland and facilitate prompt modifications to economic policies and tactics.

References

Ahmed, A., Khan, S., & Liaqat, I. (2022). Machine learning approaches to forecast the Consumer Price Index: An application to developing countries. *Journal of Economic Modeling, 38*(2), 256-270.

Brock, W. A., Dechert, W. D., Scheinkman, J. A., & LeBaron, B. (1987). A test for independence based on the correlation dimension. *Econometric Reviews, 15*(3), 197-235.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W. & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 1(1), 3146-3154.

Li, X., Feng, Y., & Xu, C. (2019). Predicting consumer price index based on deep learning. *Computational Economics, 54*(3), 987-1007.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLOS ONE, 13*(3), 29-48.

Mohamed, J. (2020). Time series modeling and forecasting of Somaliland consumer price index: A comparison of ARIMA and regression with ARIMA errors. *American Journal of Theoretical and Applied Statistics*, *9*(4), 143-153.

Van, L. T. H., & Bao, H. D. (2019). Forecasting Consumer Price Index with machine learning techniques: Evidence from Vietnam. *Asian Economic and Financial Review*, *9*(4), 460-474.

Zhang, L., Liu, W., & Wang, Y. (2020). A hybrid approach for CPI prediction based on ARIMA and XGBoost. *Applied Economics Letters*, *27*(16), 1291-1295.