MODELLING THE NAIRA/U.S. DOLLAR CURRENCY EXCHANGE RATES USING DECISION TREE, ORDINARY LEAST SQUARES AND RANDOM FOREST MACHINE LEARNING ALGORITHMS

*** [1]IBEKWE, U. A. & [2]AJIJOLA, L. A.**
[1,2]Department of Actuarial Science & Insurance
Faculty of Management Sciences, University of Lagos. Nigeria.
*ucheibek@yayoo.com
***Correspondence**

Abstract

The exchange rate of a country's currency is the fundamental price in any economy and provides a measure of the value of that country's currency relative to other currencies. This paper examines the issues of model fit and prediction accuracy of three modelling techniques that could be employed in analyzing historical data regarding currency exchange rates of the Naira with the U.S. Dollar as a reference. The techniques of Ordinary Least Squares (OLS), Decision Trees (DTs), and Random Forest were analyzed and their performance was compared in terms of model fit and prediction accuracy. The study found that Random Forest offered the best results, followed by Ordinary Least Squares and then Decision Trees, in that order. The implication and recommendation, therefore, is that the Random Forest model should be preferred in future studies in this area, although OLS also gives reasonable results and could also be deployed.

Keywords: Currency Exchange Rates, Decision Tree, Machine Learning, OLS, Random Forest.

1.0 Introduction

The exchange rate of a country's currency is a very important variable in the socioeconomic policy-making of any country (Oke, 2015). The currency exchange rate is an important measure of the economic health of a country, and affects the country's balance of trade, and influences interest rates and inflation (Rodrigues et al., 2020). It also has a major impact on international investment flows. A currency's exchange rate has been described as the fundamental price in an economy. This price gives an indication of the value of the country's currency relative to other currencies (Levinson, 2005).

A good grasp or knowledge of the models suitable for predicting the probable trajectory of a country's exchange rates is fundamentally important in providing useful information regarding the economy. According to Rahayu et al. (2017), exchange rate determination involves the concepts of purchasing power parity and balance of payments theory. Owoeye and Ogunmakin (2013) explained that exchange rate directly affects the domestic price level, the profitability of traded goods and services, as well as allocation of resources in an economy.

Currencies are usually traded in four separates but closely related markets namely, the spot market, the futures market, the options market and the derivatives market. The participants in the foreign exchange markets include exporters and importers, investors, speculators and governments (Levinson, 2005). The currency markets do not have any particular physical location, but most trading is computerized or done by telephone among financial institutions and interbank markets.

The main risk in currency trading arises from trading across different time zones. Sovereign risk is also of prime importance. As explained in Bordo et al. (2009), exchange rate depreciation can increase sovereign risk if a country has to repay its debts in foreign currency even as the local currency sheds value. Exchange rates can be quite volatile, particularly in the short term. In the long run, expectations of real interest rates drive exchange rates. This is achieved through the mechanism of covered interest arbitrage. The reason foreign exchange rate can be quite volatile is because it is largely influenced by the psychology of market participants which is usually based on economic and political factors (Zahrah et al. 2021).

Governments manage exchange rates through one of three approaches: fixed, semi-fixed and floating exchange rate regimes. Over the years the Nigerian currency has experienced all three regimes (Oke et al., 2015). Fixed rate regimes include the Gold Standard, Bretton woods (based on gold and foreign currencies), and Pegs. In pegged rates, the country keeps the value of its currency constant in terms of another currency. One major problem of fixed rate systems is the lack of flexibility regarding monetary policy options. Central banks under fixed rate systems are thus constrained to devote monetary policy to the sole aim of maintaining a stable exchange rate, to the exclusion of other goals like fighting inflation or attempting to revive a depressed economy.

Semi-fixed or hybrid systems result from attempts to provide more flexible monetary options, enabling governments to pursue other economic goals in addition to exchange rate stability.

In floating rate systems, exchange rates are determined by market forces, freeing the governments and central banks to pursue other goals. In 1986, the Nigerian currency was liberalized and opened to market forces through the Structural Adjustment Programme (SAP) introduced by the IMF.

The exchange rate is an important determinant of the economic development of a country. This study, therefore examines some suitable models for the determination of the Naira/US Dollar exchange rates. The study explores three modelling techniques: Ordinary Least Squares (OLS), Decision Tree (DT), and Random Forest (RF). From the econometric models' approach, OLS is selected because it is one of the basic tools used in econometrics for investigating or determining relationships among a set of variables. Having determined the relationships among the variables of interest, a study is expected to provide guidance for policy decisions. Decision trees are a robust tool in data science and other quantitative disciplines for decision-making. Because decision trees, more often than not, tend to involve issues of overfitting, Random Forest technique is employed here to address the issues surrounding decision trees. The primary objective is to assess the performance of these contemporary methods (OLS, Decision Tree, and Random Forest) for model fit and prediction accuracy in modelling the evolution of the Naira/USD exchange rates using historical data. This study does not set out to estimate or forecast exchange rates. Hence the various methods and tools of forecasting are not treated here. Also, the selected models were chosen for ease of coding. Other models can be explored in later research.

2.0 Literature Review

2.1 Theoretical Review

There are several approaches to currency exchange rate determination. These include the traditional approach, the Purchasing Power Parity (PPP) approach, the relative economic strength approach and the econometric models of forecasting exchange rates. The traditional approach acknowledges the role of the forces of demand and supply; the purchasing power parity approach emphasizes the law of one price in economics whereby identical goods in different should have identical prices; the relative economic strength approach assesses and compares the strength of economic growth in different countries, while the econometric models formulate models to explain the relationships among the variables of interest and make forecasts.

2.2 Empirical Review

Various empirical studies have been done on the determinants of foreign exchange rates in Nigeria and other developing countries.

Ibekwe and Shiro (2022) modelled the determinants of Naira/US Dollar currency exchange rates using Principal Component Analysis (PCA) and Singular Value Decomposition (SVD). They investigated the twin problems of multicollinearity and redundancy in explanatory variables. The researchers deployed Machine Learning algorithms to achieve dimensionality reduction thereby resolving the issues of multicollinearity and high dimensionality. In addition, the study found that three principal components adequately explain more than seventy percent of the variance while five principal components explain more than ninety percent of the variance, thus recommending the use of between three and five explanatory variables in similar studies.

Zahrah et al. (2021) implemented Long-Short Term Memory (LSTM) model in predicting foreign exchange rates to obtain the best hyperparameters based on RMSE evaluation. The best hyperparameters in 1-hour timeframe were found to be 1 hidden layer and five neurons without a dropout layer. The results for 2020 were then compared with those of 2018 and 2019.

Rodrigues et al. (2020) analyzed the behavior of currency exchange rates with Singular Spectrum Analysis (SSA) and Artificial Neural Networks (ANN). The authors found that the best performance was obtained by a hybrid method combining SSA and ANN.

Islam and Hossain (2020) presented a new model that combined Gated Recurrent Unit (GRU) and Long-Short term Memory (LSTM) for forecasting the prices of foreign exchange currencies. Results showed that the hybrid model predicted the prices of foreign currencies more accurately than standalone GRU or LSTM.

Qu and Zhao (2019) applied LSTM Neural Network models in Deep Learning to predict the price of foreign exchange. Using technical analysis, they compared two deep learning models - LSTM Neural Network and RNN Neural Network. Their results indicated that the LSTM outperformed the RNN Network model.

Oke and Adetan (2018) empirically analyzed the determinants of exchange rate in Nigeria using the ARDL Bounds test co-integration approach for the period 1986-2016. They found that the GDP, Interest rate and Inflation rate positively affected the exchange rate while Degree of Openness had negative effect.

Rahayu et al. (2017) used Principal Component Analysis to reduce multicollinearity in the factors affecting currency exchange rates of some Asian countries. They used multiple regression to examine the differences in multicollinearity at yield. Results indicated that the proportion of variance explained by one principal component was as high as ninety-eight percent.

Ajao (2015) in Oke (2018) examined the factors that determine real exchange rate volatility in Nigeria. Using GARCH (1,1) and the Error Correction Model (ECM) the author carried out co-integration analysis and found that lagged exchange rate, government expenditures, interest rate movements, and openness of the economy were key determinants of volatility in real exchange rates.

Regos (2015) modelled the exchange rate using price levels and country risk. The study built two factor discrete time models in a Markov switching framework to investigate the effect of sovereign risk on the nominal exchange rate. The study found that sovereign risk has a significant effect on nominal exchange rates.

Alayande (2014) employed the unit root test and granger causality test to investigate the factors that affect the exchange rate in Nigeria using data from 1980 to 2013. The study revealed that growth in money supply, foreign exchange reserves, interest rates, and the rate of inflation affect exchange rates in Nigeria.

3.0 Materials and Methods

3.1 Data

This paper uses monthly data on the average Naira /US Dollar exchange rates and other relevant macroeconomic variables including interest rates, external reserves, balance of payments, crude oil price etc., from the Central Bank of Nigeria database over the period 2008 to 2021 inclusive.

The variables in this study are drawn from the literature regarding the factors that determine currency exchange rates. These include Balance of payments (Rahayu et a., 2017), Government expenditure, interest rates, inflation rates (Oke et al., 2015), import cover, exports, external reserves, deposit and lending rates (Ibekwe and Shiro, 2022).

3.2 Methods

Exploratory data analysis is carried out using Machine Learning Python codes. Three different models-Ordinary Least Squares regression (OLS), Decision Trees (DTs) and Random Forest Machine Learning Algorithms-were built and deployed to analyze time series historical data. The performances of the models are compared in terms of model fit and prediction accuracy using Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). Time series regression is not used here because, as explained by Montgomery and Hines (1990) in Rahayu et al. (2017), the impact of multicollinearities can result in regression coefficients becoming rather weak. More importantly, the focus of this study is neither forecasting nor causation.

3.2.1 Ordinary Least Squares (OLS)

Ordinary Least Squares regression is a statistical method for estimating the coefficients of linear regression equations. OLS indicates the relationship between a target variable and one or more explanatory variables. OLS estimates the unknown parameters in the regression model by using the method of least squares to minimize the sum of the squares of the differences between the values of the variable being observed and the values being predicted. The smaller the differences between the observed and the predicted, the better the model fits the data. OLS provides mean-unbiased and minimum-variance estimation if the errors are homoscedastic and serially uncorrelated. Other assumptions to guarantee the validity of OLS for estimating the regression coefficients include linearity, random sampling and exogeneity (Sluijmers, 2020).

Given x_i, y_i
Calculate $\bar{x}, \bar{y}$ and do $x_i - \bar{x};\ y_i - \bar{y}$
Calculate or multiply $(x_i - \bar{x})(\ y_i - \bar{y})$ and sum Σ
Calculate $(x_i - \bar{x})^2$
Calculate the slope = $\frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$ (3.1)
The slope and y-intercept are used to form the line of best fit
Y = slope(x) + intercept
Draw on scatter plot

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 x + \beta_2 x_2 + \cdots + \beta_i x_i \qquad (3.2)$$

Where Y = Dependent variable

Regression-Evaluation Method

Some regression evaluation metrics include R-squared; Adjusted R-squared; Mean Absolute Error; Root Mean Square Error (Songhao, 2020).
R^2: This is a measure of the percentage of variance in the dependent variable that is explained by the model. Usually, this is the first metric employed for assessing linear regression model performance. The higher the value of R-squared, the better.

$$R^2 = 1 - \frac{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - \bar{y})^2} \qquad (3.3)$$

Adjusted R-squared: Conceptually, this metric is like R-squared, but is adversely affected If there are too many variables. Generally R-squared increases as the number of variables increases, but this does not affect adjusted R^2. The higher the value of adjusted R-squared, better.

$$Adjusted\ R^2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1} \qquad (3.4)$$

Where p = Number of independent variables

Mean Absolute Error (MAE): This is the simplest metric for assessing prediction accuracy. It has the same units as the target variable. It is not sensitive to outliers.The lower the value of MAE, the better.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (3.5)$$

Root Mean Square Error (RMSE): This is another metric used to assess prediction accuracy. It has the same units as the explanatory variable. Because it is sensitive to outliers, errors will be magnified as a result of the square function.The lower the value of RMSE, the better.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (3.6)$$

Bias: This is the difference between the prediction of the model and the actual value.

Variance: This is an indication of the spread of the data.

Bias and Variance are used to describe whether the model is underfitting or overfitting.

3.3 CROSS-VALIDATION

This is a statistical method employed to address the issue of overfitting in a predictive model. Cross-validation is used to estimate the performance of Machine Learning models in terms of their accuracy. Cross-validation is a technique for assessing how generalizable a model is to unseen data (Shayan et al., 2018).

In Cross-validation, the training data is repeatedly split or partitioned into a fixed number of folds to estimate the potential out-of-sample predictive performance. The data in each fold is analyzed, and the error estimates from the different folds are averaged. This proceeds in 3 steps:

Step 1: Split the data into k partitions or folds.

Step 2: Leave one fold out for testing while a prediction rule is adopted for all other folds.Then predict outcome observations in the fold that was left out, and record the empirical Mean Squared Prediction Error. Repeat this process for each fold.

Step 3: Average the empirical Mean Squared Prediction Errors over all the folds.

Repeat these steps for several values of the tuning parameters and choose the best tuning parameter that minimizes Average Mean Squared Prediction Error.

3.4 Decision Trees (DTs)

Decision Trees (DTs) are a predictive modeling technique used in classification and regression. They are used to obtain the value of a target variable by making use of decision rules deduced from features of the data. In Python computer language, DecisionTreeClassifier performs multi-class classification on a dataset while for regression, DecisionTreeRegressor is used. Decision tree technique is a data mining method employed in classification and prediction systems (Song and Lu, 2015).

Note that a tree with few samples in high dimensional space is likely to overfit.

A decision tree recursively partitions the feature space in such a way that samples with same labels or similar target values are grouped together. To avoid over-fitting, the decision tree has to be pruned. An algorithm called minimal cost-complexity pruning is used to prune the tree.

3.5 Random Forest Regressor

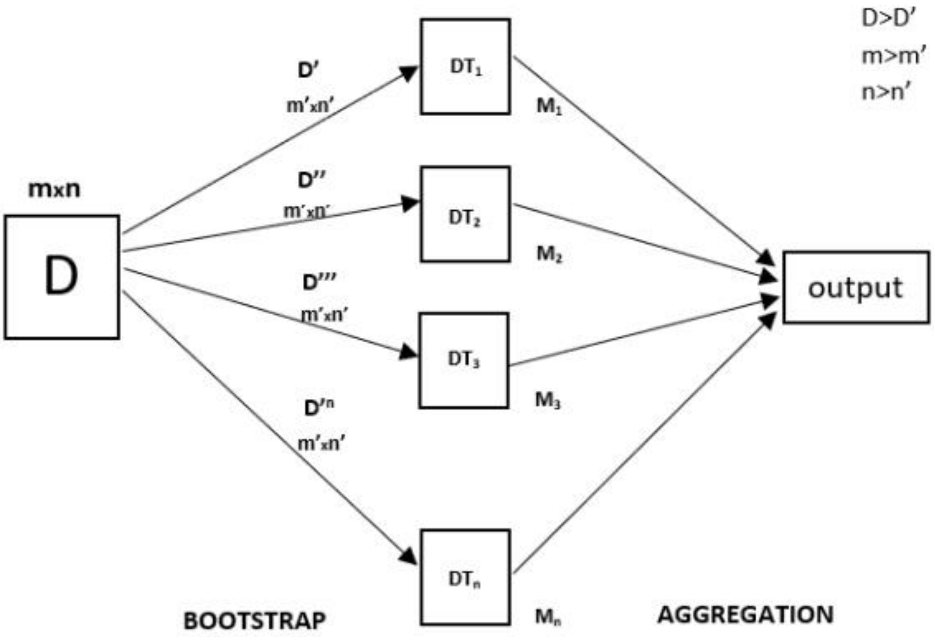
Random Forest is an estimator that deploys more than one decision tree on different sub-samples of the data set. The output of the Random Forest is the mean of the outputs of the component decision trees. Using the mean values of the sub-samples improves the predictive accuracy of the model and helps reduce overfitting.

Random Forest regression technique:

First, a machine learning model is created to train the data and set the baseline model. Then obtain insights from the model using the test data. After this, the performance metrics of the test and the predicted data are compared. Try improving your model using another data modeling technique. Interpret and report your observations.

Random Forest is a statistical tool employed in performing regression and classification tasks using several or many decision trees and a technique called Bootstrap and Aggregation, also known as bagging. Feature and row sampling are performed on the data set to obtain sample datasets for every model. This is called Bootstrap.

Fig. 1 Bootstrap Aggregation



Source: GeeksforGeeks.org

In regression, the mean prediction of the trees is the output, while in classification it is the mode. Random forests adequately address the issue of overfitting. Random forest algorithm uses bootstrap aggregating or bagging.

Bootstrapping results in improved model performance. It reduces model variance without increasing the bias.

Implementation of Random Forest Regression in Python**.**

Step 1: Import the required libraries.
Step 2: Import the data set
Step 3: Assign all rows and the first column to x, and all rows and the second column to y
Step 4: Fit Random forest regressor to the dataset Output
Step 5: Predict a new result
Step 6: Visualize the result Output**:**

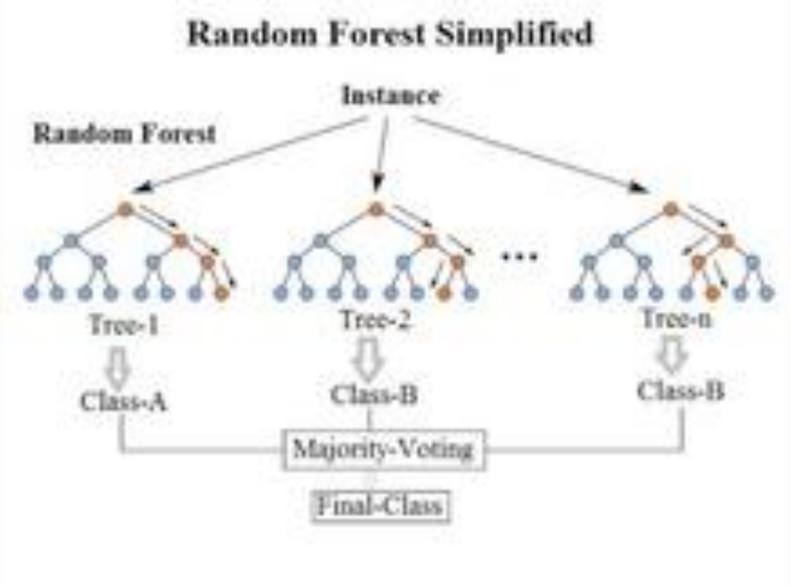


Fig. 2 Random Forest

Diagram of a random decision forest (Wikipedia).

Each decision tree normally has high variance, but when combined in parallel the overall variance is low since each decision tree is perfectly trained on its very own sub-sample. The output depends on the output of more than one tree. In classification problems, the final output is the majority among voting classifiers. In regression, the final output is the average value of all the outputs. This is called Aggregation. Random forest is an ensemble classification method of high accuracy (Fawagreh et al., 2014).

4.0 Results and Discussion

The dataset contains 1,848 data points arranged in 168 rows and 11 columns. All the columns are numeric and have non-null values.

Table 1. Summary statistics

	count	mean	std	min	25%	50%	75%	max
Foreign Reserves Position	168	38669.34	8188.968	23689.87	32971.78	37130.23	43367.09	62081.86
Crude Oil Price (US$/Barrel)	168	77.60018	27.5435	14.28	55.385	73.555	106.195	138.74
Savings Rate	168	2.912083	0.984059	1.25	1.83	3.135	3.89	4.28
Int Rate CmBnk Tdep 1 Yr	168	8.275536	2.842299	3.53	5.64	8.015	10.575	16.47
Prime Lending Rate CmBnk	168	16.09268	2.029009	11.13	15.505	16.595	17.1025	19.66
Max Lending Rate CmBnks	168	25.96018	3.651563	17.58	23.195	26.07	28.74	31.56
Imports (CIF)	168	4470.279	1231.954	2502.6	3521.625	4441.59	5175.183	8574.64
Balance of Trade	168	1348.08	1657.501	-3480.64	0	1050.955	2356.935	5996.98
All items>Y-o-Y change (%)	168	12.31685	3.021637	7.71	9.5875	12.065	14.445	18.72
Food>Y-o-Y change (%)	168	13.97048	3.770191	7.88	10.19	13.46	16.63	22.95
Average Exchange Rate	168	226.4696	90.93557	117.72	153.7925	169.68	306.28	414.34

UNIVARIATE ANALYSIS

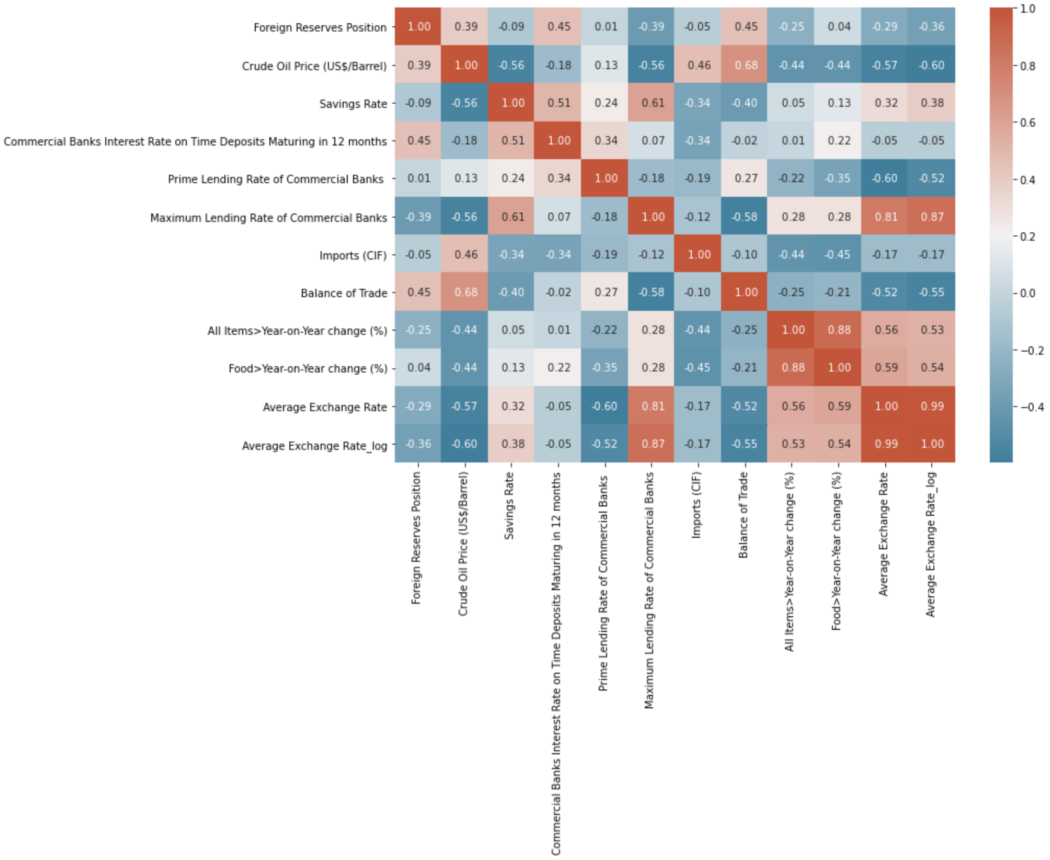
Observations:

The distributions of the variables were checked for outliers. Distribution plots showed that Imports (CIF), Balance of Trade and Foreign Reserves Position are highly right skewed.The target variable, Average Exchange Rate is not normally distributed. Therefore we do a log transformation.

CORRELATIONS

Using the heat map, we find few correlations among the variables, with the exception of moderate positive correlation between crude oil price and balance of trade, as expected. This means that the problem of multicollinearity was largely addressed in the choice of explanatory variables.

Fig. 3 Heat map



Source: sns.heatmap output

4.1 Model Building

This involves the following steps:

1. Data preparation
2. Partitioning the data into train and test sets
3. Building model on the train data
4. Cross validating the model
5. Testing the data on test set.

SPLIT THE DATASET

Having prepared the data in previous sections, the data is now split into dependent and independent variables, and further divided into train set and test set in a ratio of 70:30. The first five rows of the train set is shown below. In python computer language this is obtained as X_train.head().

Table 2. X_train.head()

69	1	44155.11	112.29	2.39	4.71	17.1	24.9	5547.4	1919.76	7.8	9.2
28	1	38815.79	77.5	2.92	6.31	18.77	22.56	4400.7	2164.19	12.91	13.02
58	1	42568.26	111.05	1.65	6.17	16.51	24.7	4639.2	3814.33	12.32	11.55
153	1	35580.48	39.74	1.87	4.99	11.31	28.36	6434.2	0	14.23	17.38
66	1	45834.11	109.78	2.42	5.8	16.47	24.62	4968.7	3498.21	8.68	9.99

Next check for multicollinearity by calculating VIF for each feature.

Table 3. VIF for each feature

	Feature			Feature
0	450.57		5	2.6159
1	3.2354		6	2.6652
2	4.6887		7	2.7435
3	3.345		8	3.827
4	2.9588		9	11.248
			10	12.566

Observations: Most of the values are less than 10
Recreate the model after dropping feature number 10 (Food>Year-on-Year change (%)) which had the highest VIF value of 12.566 > 10.

Table 4. VIF after dropping feature number 10

	Feature		Feature
0	448.99	5	1.9664
1	2.8449	6	2.6352
2	4.5907	7	2.742
3	3.3421	8	3.7049
4	2.5006	9	1.7241

Observations:
Now VIF < 5 for all variables.
Next step, create linear regression model using stas model OLS.
4.2 Create the model. Get the model summary.
model1.summary()

Tables 5 and 6 are standard Data Science and Machine Learning Python outputs for OLS. The most important columns are referenced and explained under the relevant segment immediately following each table and titled 'Observations'.

4.2 OLS Regression Results:

Table 5. **model1.summary ()**

OLS Regression Results

Dep. Variable:	Average Exchange Rate_log	**R-squared:**	0.954
Model:	OLS	**Adj. R-squared:**	0.949
Method:	Least Squares	**F-statistic:**	218.9
Date:	Fri, 03 Jun 2022	**Prob (F-statistic):**	5.44e-66
Time:	14:38:29	**Log-Likelihood:**	124.86
No. Observations:	117	**AIC:**	-227.7
Df Residuals:	106	**BIC:**	-197.3
Df Model:	10		
Covariance Type:	nonrobust		

coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**	
const	4.1469	0.172	24.166	0.000	3.807	4.487
Foreign Reserves Position	-2.295e-07	1.75e-06	-0.131	0.896	-3.7e-06	3.24e-06
Crude Oil Price (US$/Barrel)	-0.0020	0.001	-3.158	0.002	-0.003	-0.001
Savings Rate	-0.0110	0.015	-0.715	0.476	-0.042	0.019
Commercial Banks Interest Rate on Time Deposits Maturing in 12 months	0.0068	0.005	1.397	0.165	-0.003	0.016
Prime Lending Rate of Commercial Banks	-0.0688	0.006	-10.647	0.000	-0.082	-0.056
Maximum Lending Rate of Commercial Banks	0.0793	0.004	22.193	0.000	0.072	0.086
Imports (CIF)	7.673e-06	1.08e-05	0.712	0.478	-1.37e-	2.9e-05

					05	
Balance of Trade	2.949e-05	9.38e-06	3.143	0.002	1.09e-05	4.81e-05
All Items>Year-on-Year change (%)	0.0396	0.009	4.446	0.000	0.022	0.057
Food>Year-on-Year change (%)	-0.0121	0.008	-1.607	0.111	-0.027	0.003

Omnibus:	8.038	**Durbin-Watson:**	2.124
Prob(Omnibus):	0.018	**Jarque-Bera (JB):**	16.275
Skew:	0.009	**Prob(JB):**	0.000292
Kurtosis:	4.827	**Cond. No.**	8.57e+05

Notes:
[1] The covariance matrix of the errors must be correctly specified.
[2] A large condition number could be indicative of
strong multicollinearity or other numerical problems.

Observations: For independent variables, the lower the std error, the better. P-value is used to determine whether the variables have significant relationship.
P > 0.05 means no significant relationship; p < 0.05 significant: Significance means the population regression parameters are significantly different from zero.
Eliminate those > 0.05 and create a new model.
Examine the significance of the model. Check whether all the coefficients are significant. Remove those > 0.05
Create the model after dropping Food>Year-on-Year change (%).
Split the data in 70:30 ratio of train to test data, create the model and get the model summary.

4.3 OLS Regression Results:

Table 6. **model2.summary()**

Dep. Variable:	Average Exchange Rate_log	**R-squared:**	0.953
Model:	OLS	**Adj. R-squared:**	0.949
Method:	Least Squares	**F-statistic:**	239.4
Date:	Tue, 17 May 2022	**Prob (F-statistic):**	1.22E-66
Time:	17:33:26	**Log-**	123.45

		Likelihood:				
No. Observations:	117	**AIC:**	-226.9			
Df Residuals:	107	**BIC:**	-199.3			
Df Model:	9					
Covariance Type:	nonrobust					
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	4.1306	0.173	23.937	0	3.788	4.473
Foreign Reserves Position	-1.21E-06	1.65E-06	-0.73	0.47	-4.49E-06	2.07E-06
Crude Oil Price (US$/Barrel)	-0.0019	0.001	-2.935	0	-0.003	-0.001
Savings Rate	-0.0117	0.015	-0.758	0.45	-0.042	0.019
Int Rate CmBnk Tdep 1 Yr	0.0037	0.004	0.826	0.41	-0.005	0.013
Prime Lending Rate CmBnk	-0.0637	0.006	-11.274	0	-0.075	-0.052
Max Lending Rate CmBnks	0.0787	0.004	21.985	0	0.072	0.086
Imports (CIF)	7.27E-06	1.09E-05	0.67	0.5	-1.42E-05	2.88E-05
Balance of Trade	2.68E-05	9.30E-06	2.881	0.01	8.36E-06	4.52E-05
All items>Y-o-Y change (%)	0.0264	0.004	7.525	0	0.019	0.033
Omnibus:	7.389	**Durbin-Watson:**	2.149			
Prob(Omnibus):	0.025	**Jarque-Bera (JB):**	10.938			
Skew:	0.258	**Prob(JB):**	0.00421			
Kurtosis:	4.406	**Cond. No.**	8.56E+05			

Notes:

[1] It is important to ensure that the covariance matrix of the errors is correctly specified.

[2] The large condition number, 8.56e+05, might indicate strong multicollinearity or other numerical problems.

Observations: For independent variables, the lower the std error, the better.
P-value is used to determine whether the variables have significant relationship.

Observations

Checking the performance of the model on the train and test dataset

targets = vol['Average Exchange Rate']

Table 7. targets.head ()

```
0    117.98
1    118.21
2    117.92
3    117.87
4    117.83
Name: Average Exchange Rate
```

Check model performance

Table 8. Model performance

	Data	RMSE	MAE	MAPE
0	Train	0.084241	0.062685	1.185207
1	Test	0.09037	0.06552	1.221432

Observations: RMSE,MAE and MAPE of train and test data are not very different, indicating that the model is not overfitting and has generalized well.

Apply the cross validation technique to improve the model and evaluate it using different evaluation metrics.

Import the required function
Build the regression model using sklearn linear
Regression.

Divide Data into 10 folds
We get 10 rsquared values, 0.729 is the average with a range of+/-0.232 (tolerance level)
The tolerance level gives an idea how robust the model is.If robustness is high, tolerance level will be low.
RSquared:0.940(+/-0.057)
MeanSquaredError:0.009(+/-0.011)

Observations: The mean squared error is good. The higher, the value, the better.
Get model coefficients in a pandas data frame with column "Feature" having all the features and column 'coefs' with all the corresponding coefficients. Write the regression equation using
coef=model2.params

Table 9. Model 2 parameters

const	4.130582
Foreign Reserves Position	-0.0000010
Crude Oil Price (US$/Barrel)	-0.001899
Savings Rate	-0.011743
ComBnks Int. Rates on 12 months T Deps	0.003706
Prime Lending Rate of Commercial Banks	-0.063665
Max Lending Rates of Commercial Banks	0.07869
Imports (CIF)	0.000007
Balance of Trade	0.000027
All Items>Year-on-Year change (%)	0.026399

Equation of the fit
log(Average Exchange Rate) = (4.130582268602491)* const +(-1.207221938679842e-06)* Foreign Reserves Position +(-0.0018987413845580926)* Crude Oil Price (US$/Barrel) +(-0.011742845874719045)* Savings Rate +(0.0037062303403001 54)* CBIRTD 12-month Maturity +(-0.06366508331794007)* PLRCB +(0.07868983853409803)* MLRCB +(7.269066344809462e-06)* Imports (CIF) +(2.6796227063352853e-05)* Balance of Trade +(0.02639926662693559)* All Items>Year-on-Year change (%)

Observations:
Interpreting Regression Coefficients involves holding all other dependent variables constant while changing the value of one explanatory variable by one unit and observing how the target variable changes in response to the unit change.

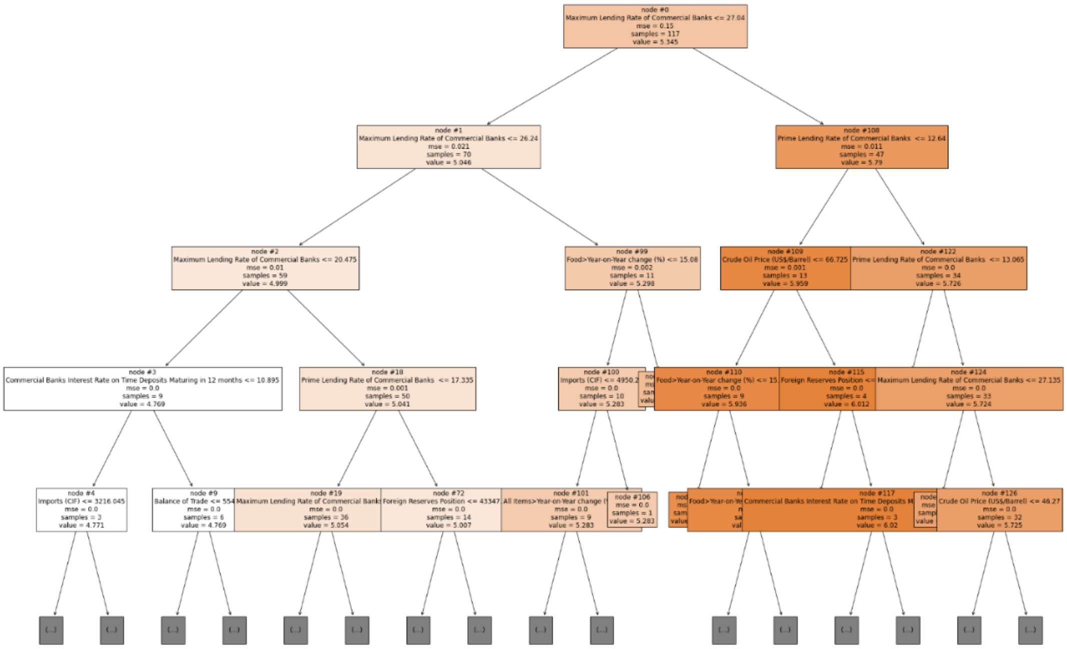
4.4 Building the Non-Linear models (Decision Tree and Random Forest) and checking their performance. This involves:
# Importing Decision Tree Regressor using sklearn
# Splitting the data in 70:30 ratio of train to test data.
# Separating the dependent and independent variables
# Redefining the Decision tree regressor
# Fitting Decision Tree Regressor to train dataset
# Checking model performance on the train and test set

Table 10. Model Performance

	Data	RMSE	MAE	MAPE
0	Train	2.172483E-16	5.313888E-17	8.941725E-16
1	Test	2.796369E-02	1.021510E-02	1.991520E-01

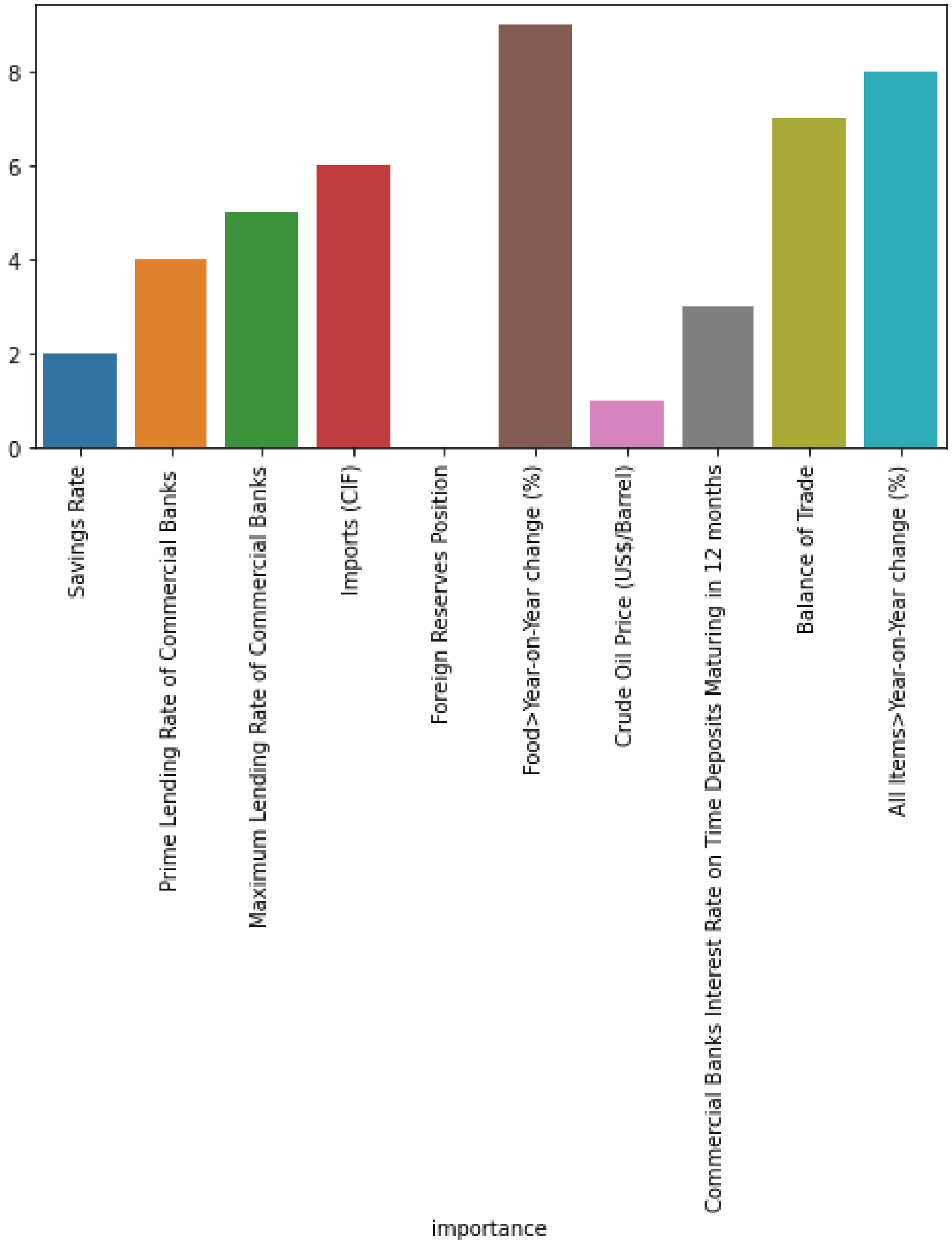
Observations: Test set values are of much higher order than the train set. The model is clearly overfitting and does not generalize well.
Decision trees tend to have high variance, but they are not difficult to interpret. Splitting is based on maximum information gain. It is based on the variable that gives least variance.

Fig. 4 Decision Tree



Observations: The tree is of depth five. Greater depth is required for more complex relationships.
Plot the feature importance for each variable in the dataset and analyze the variables.

Fig. 5 Feature importance for each variable.



Source: sns.barplot output

4.5 Building Random Forest: In building Random Forest, hyperparameters are randomly chosen initially but adjustments are subsequently made to obtain the best model. Model building involves:

# Splitting the model.
# Checking model performance on the train and test dataset.
# Score gives the accuracy of the test data and labels. Here, the score was 0.9961373242071023
# Observations: The higher the score the better.

Table 11. Model performance on train and test set

	Data	RMSE	MAE	MAPE
0	Train	0.018921	0.008753	0.16424
1	Test	0.024303	0.013355	0.247556

Observations: RMSE,MAE and MAPE for the random forest are small and are close for both train and test dataset. Hence, the model is performing very well and giving generalized results.

5.0 Conclusion

This study compared three modelling techniques that could be used to model the evolution of the Naira/US Dollar currency exchange rates. The models were Decision Tree, OLS and Random Forest machine learning algorithms.The three models techniques were compared in terms of model fit and prediction accuracy. Model performance was measured using RMSE, MAE and MAPE.

The study found that OLS results for train and test data were not very different, indicating that the model generalized well and was not overfitting.

The Decision Tree model, however, was high in variance and did not generalize well. The DT model was highly overfitting as the order of differences between train and test data using RMSE, MAE and MAPE was rather large.

Random Forest algorithm returned very high accuracy level (99.6 percent) on test data and labels. RMSE, MAE and MAPE for the random forest are small and are close for both train and test data set. Hence, the model is performing very well and giving generalized results.

Clearly, Random Forest offered the best results, followed by OLS and then Decision Tree models in that order. The implication and recommendation is therefore that Random Forest model should be deployed in future studies in this area of research, but OLS also gives reasonable results.

REFERENCES

Ajao, M.G. (2015).The Determinants of Real Exchange Rate Volatility in Nige-ria. Ethiopian Journal of Economics, *24* (2), 44-62.

Alayande, S.A.(2014). Modelling Exchange Rate in Nigeria. *International Journal of Research in Applied Natural and Social Sciences, 2*(6), 169-176.

Bordo,M.D., Meissner, C.M., and Weidenmier,M. D. (2009). Identifying the Effects of an Exchange Rate Depreciation on Country Risk. Evidence from a Natural Experiment. *Journal of International Money and Finance, 28,*1022-1044.

Fawagreh,K., Gaber, M.M. and Elyan, E.Random Forests: From Early Developments to Recent Advancements. *Systems Science and Control Engineering, 2*(1), 602-609

Geekforgeeks.org (2022). *Computer Science Portal for Geeks* https://geekforgeeks.org.

Ibekwe, U.A. and Shiro, A.A. (2022). Modelling the Determinants of Naira/U.S. Dollar Currency Exchange Rates Using Principal Component Analysis (PCA) and Singular Value Decomposition (SVD). *Nigeria Journal of Risk and Insurance, 12* (1)2022, 77-97. *http://njri.unilag.edu.ng*

Islam, M.S. and Hossain, E. (2020). Foreign Exchange Currency Rate Prediction Using a GRU-LSTM hybrid Network.*Soft Computing Letters* *www.elsevier.com/locate/socl,https://doi.org/10.1016/j.socl.2020.100009*

Levinson, M.(2005). *The Economist Guide to Financial Markets,Fourth Edition*, p.14-15.

Oke, M.O. and Adetan, T.T.(2018). An Empirical Analysis of the Determinants of Exchange Rate in Nigeria. *International Journal of Scientific Research and Management (IJSRM), 6, (*5).

Owoeye,T. and Ogunmakin, A.A. (2013). Exchange Rate Volatility and Bank Performance in Nigeria. *Asian Economic and Financial Review,3*(2),178-185.

Qu, Y. and Zhao,X.(2019). Application of LSTM Neural Network in Forecasting Foreign Exchange Price. *Journal of Physics: Conference Series.1237 042036.*

Rahayu,S., Sugiarto,T., Madu,L., Holiawati, and Subagyo, A.. Application of Principal Component Analysis (PCA) to Reduce Multicollinearity Exchange Rate Currency of Some Asian Countries, Period 2004-2014. *Internat-ional Journal of Educational Methodology,3*(2),75-83. *https://doi.org/10.12 973/ijem.3.2.75.*

Regos,G.(2015). Modeling the Exchange Rate Using Price Levels and Country Risk. *Cogent Economics & Finance, 3 (*1)*, 1056928, DOI:10.1080/23322039.2015.1056928*

Rodrigues, P.C., Awe, O. O. Pimentel, J.S., and Mahmoudv and, R. (2021). M odelling the Behaviour of Currency Exchange Rates with Singular Spect rum Analysis and doi*10.3390/stats3020012*www.mdpi.com/journal/stats.

Shayan, T.B., Amin, E., Sihai,D.Z. and Saurabh, S.(2018). A Closer Look at Cross-validation or Assessing the Accuracy of Gene Regulatory Networ ks and Models. *Scientific Reports*, *8*, 6620

Sluijmers, M. (2020). Simple Linear Regression and OLS: Introduction to the Theory.*https://towardsdatascience.com/simple-linear-regression-and-ol s-introduction-to-the-theory-1b48f7c69867*

Song, Y. and Lu, Y.(2015. Decision Tree Methods: Applications for Classific ation and Prediction. Shanghai Archives of Psychiatry, 27 (2), 130-135, doi:10.11919/j.issn.1002-0829.215044

Songhao, W. (2020). Three Best Metrics to evaluate Regression Model? *https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755*

Wikipedia (2022). Random Forest Simplified. https://en.wi kipedia.org/wiki/ Random_forest

Zahrah, H.H., Sa’adah, S., Rismala, R. (2021). Foreign Exchange Rate Predict ion Using Long-Short Term Memory: A Case Study in COVID-19 Pand emic. *International Journal on ICT, 6* (2), Dec 2020. pp94-105*, doi:10.2 1108/IJOICT.2020.62.538*