Developing a Hybrid ARIMA-XGBOOST Model for Analysing Mobile Money Transaction Data in Kenya.

Abstract

Introduction: This study formulated a hybrid model of ARIMA and XGBOOST models using the dataset from the Central Bank of Kenya and the objective was to formulate a Hybrid ARIMA-XGBOOST models to capture the patterns and dynamics in Mobile Money Transactions in Kenya. The study was motivated by the gaps that existed from the previous studies which lacked the ability to model both linear and non-linear trends in the data across time period. Since 2007 when Mpesa was invented, the rate of Mobile Money transactions has rose steadily and this required a complex model that will provide a clear trend.

Research Methods: To accurately model the dynamics of Mobile Monye Transactions in Kenya, this study presented a complex Hybrid model. The model was formulated by combining Autoregressive Integrated Moving Averages (ARIMA) and Extreme Gradient Boosting (XGBOOST). The ARIMA captures the linear trends of the data while the XGBOOST models the non-linear part and trained to model the ARIMA residuals. The model parameters were evaluated using Mean Absolute Error, Root Mean Square Error among others and confirmed to be accurate and reliable.

Analysis and Results: Based on the findings from the ADF coefficient, the stationary condition was met (P-value=0.01), and therefore we proceeded to develop the ARIMA models. Initial diagnoses included model identification and examination of autocorrelation to determine the ARIMA configurations, whereas the Box-Jenkins test confirmed the models' adequacy (P-value=2.220e-16). Based on Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Percentage Error (MPE), which are all below 1%, indicates that the performance of the models had high prediction accuracy.

Conclusion: Unlike other models, the Hybrid ARIMA-XGBOOST model was effective in capturing the mobile money transaction in Kenya, and the model also provided adequate forecasts. This model could greatly contribute to the endeavour of strategic planning and budgeting, replacing the shortcomings of today's forecasts and improving business, and individual's transaction forecasting capabilities.

Key Words: Mpesa, Mobile Money Transaction, Cash in Cash out Volume, Value, ARIMA, XGBOOST, Hybrid Model.

Background of the Study

In the new era of Kenyan Financial Arena, mobile money services create a new standard, the revolutionary standard which moves the traditional transaction lineage forward and bring it closer to everyone. The study attempted to reveal the interwoven patterns and rules that exist in the mobile money data, bringing a complex view of its growth and development [2]. Actually, the study tried to grasp the power of hybrid modelling, using combined ARIMA and XGBOOST algorithms, to build a strong forecasting model for mobile money transactions in Kenya [8]. This approach recognized the multi-dimensional character of the data, involving linear and non-linear dependencies [7]. The stated

goals helped clarify the research universe with the goal of refining the hybrid model to accurately represent the distinctive patterns and dynamics of mobile money transactions [6]. A holistic review of all these ensures that the study provides a more refined understanding of mobile money transfers as it happens in the past, present and the future with this aim of contributing towards both conceptual understanding and real-world applications for different stakeholders in the financial sector [9].

The automation directives have brought about significant changes in the National Payment System (NPS) of Kenyan with their traction in the late 1980s and 1990s. Some of the strategic achievements included the introduction of the Automated Teller Machines (ATMs) in 1989, the Nairobi Clearing House automation in 1998 and the creation of the Kenya Electronic Payment and Settlement System (KEPSS); in operation with the Real-Time Gross Sett [5]. This evolution triggered the period of Kenya's modernized payment systems [4]. Human major turnaround was also in March 2007 with the creation of the revolutionary M-PESA mobile money service. M-PESA, which stands for mobile money in Swahili slaked an irrepressible thirst in many communities; a need for easy cash interchanges, especially from city to countryside. From that point, a mobile money revolution has begun in Kenya which had resulted in integration of numerous mobile money services and products into the daily routine of life. By July 2022, the Central Bank of Kenya reported staggering statistics: In one month, over 1.2 billion mobile money transactions amounting to over $23 billion Interrupted Question [5]. The estimated active mobile money customers had grown to over 28 million, which exceeded more than half of the Kenyan population, and about 310,000 mobile money agents provided various services [1].

The effects of mobile money transcended dealing in financial services, but financial inclusion increased by a significant margin. The level of financial inclusion died by 26.7% in 2006 depending on the Fin Access report, however by 2021, this value was raised drastically to 83.7% [2]. People change their financial behaviour towards the rise of mobile money services usage from 28% in 2009 to 81% in 2021. Simultaneously, the number of people who used two or more financial services in Kenya increased fourfold from 18% in 2006 to 75% in 2021 [3]. An increase in active mobile money customers and person-to-person payments due to COVID- 19 pandemic reflects the significance of e-payments; so, relief measures were introduced by Central Bank of Kenya. Moving forward; the CBK seeks to stabilize gains, increase customer centricity, and foster a reliable and environmentally-efficient safe hubs of financial transactions.

A brief historical perspective of the evolution of statistical modelling approaches for mobile money transactions in Kenya, this journey had its inception on the foundation of techniques such as time-series analysis [1]. After the approach of whole changing architecture of mobile money, complexity of their pattern and dynamic, there was an attempt of paradigm shift towards more complex forecasting models [2]. The implementation of ARIMA, a classic time series model, together with XGBOOST, an advanced machine learning tool, was a tactical approach of the moving mobile money in Kenya [5] [8]. A part of the historical background comprises essential dates, for instance, the automatization of Nairobi Clearing House back in 1998 or the creation of Kenya Electronic Payment and Settlement System (KEPSS) in 2005, 88 paving the way for the development of advanced modelling techniques. Digital payments providers as well as G20 can significantly transform forecasting because of the historical development it provides and the need to create the capacity to adapt and develop further; in Kenya, we had a progressive mobile money ecosystem and Kenya should be aiming to be a global leader and a provider for the rest of the world while implementing the necessary reforms [2]. Already the statistical modelling initiative should not be considered as a reaction to passing trends but as a proactive approach recognizing revolutionary development of mobile money and its fundamental role in remaking the financial world.

Hybrid Model integrated to include ARIMA and XGBOOST Machine Learning Model.

The study conducted by Zhan, deals with the problem of precise prediction of house prices, essential for the implementation of national policies on real estate, with the help of the hybrid machine learning framework [11]. Because of the research on the forecasting efficiency enhancement, new HBO models which are based on stacking, bagging, and transformer entities are offered [6]. However, these hybrid models employ Bayesian Optimization for hyperparameter tuning so that there is a more accurate prediction as well as greater stability. This work makes use of a diverse multi-source data set that contains several 1,898,175 transactions of the Hong Kong real estate market between 1996 and 2021. Therefore, the proposed hybrid models are compared with 18 benchmark models using 13 evaluation metrics and the findings illustrate HBOS models' superiority regarding house price prediction [11]. More precisely, HBOS-CATBOOST shows better results concerning RMSE when compared to HBOB-XGBOOST and HBOT-ConvLSTM, reducing the relative RMSE values by 5.11% and 25.56%

respectivel. The research significantly contributes by providing a detailed benchmark dataset, suggesting new hybrid models, revealing a multangular performance evaluation framework, and developing the housing pricing prediction field.

Fofanah's study addresses the need for more research on price prediction in the Ethiopia Commodity Exchange (ECX) by employing three machine learning algorithms: LR-Linear Regression, XGB-Extreme Gradient Boosting, and LSTM [5]. To predict upcoming price trends for two important ECX commodities namely Coffee and Sesame the study seeks to aid investors and business planning. The goal of comparative analysis of the algorithms is one of the primary ones. With accuracy in predictive performance as a court of judgment regarding algorithm efficiency, binary classifiers under assay employ datasets of size 7205 and 1540 instances for sesame and coffee, respectively. Moreover, Fofanah delivers the Ethiopia Coffee Prices Predictor (ECPP) application for Android gadgets that incorporates forecast calculations intended to function in the conversation office [3]. The programming environment consists of the prediction. The findings on the comparative analysis between LR, XGB, and LSTM show their effectiveness in predicting ECX commodity prices and ultimately justify developing an ECPP mobile application that is supposed to make such predictions convenient and user-friendly.

Therefore, conducted an extensive empirical investigation on the status and prospects of the RNN for a time series forecasting. Thus, the study will look at the emergence of RNNs to establish whether they outcompete the traditional statistical models such as ETS and ARIMA, which although precise, may be considered inconvenient for those who are not software experts because of ease, efficiency, and use of automation. The goals include giving pointers and best practices for using what already exists in terms of RNN architectures when forecasting [11]. The findings show that while RNNs, as represented by the best method in the M4 competition, are good competing techniques in most cases, they remain weak in robustness, efficiency, and automatic process compared to ETS and ARIMA. The study implies that RNNs are suitable for modelling seasonality directly for homogenous datasets, but the opposite could apply to heterogeneous datasets. Because of that, a deseasonalization step is also recommended. Overall, the results enrich the understanding of prospects and constraints of RNNs usage and suggest several recommendations for its enhanced application in time series forecasting [11].

According to Swamy and Sarojamma of 2017, it was noted that it was very critical to discover suitable models for transactions in the bank through which one could evaluate future transactions with respect to the prior known data [3]. Thus, this study aims to specify properly the following objectives: Utilize DBN and NN classifiers in simulating the bank transaction count and amount. In the research study, feature extraction is carried out to obtain the statistical features out of the data, merged and empathetic DBN-NN model; Average of the modelled output is adopted in both the prediction models for the count of transactions and amount based on the average modelled output as a final prediction. The latter contribution of the study is enhanced by a hybridized optimization model known as L-ABC Model, and the L-ABC models incorporate Lion Algorithm and Artificial Bee Colony algorithm to arrive at the optimal number of hidden neurons in both DBN and NN [3]. S In addition, information from model 1 which is transaction count prediction is transformed into an ARIMA model to establish the correlation between the total transaction count and the corresponding amount for the auto-regressive method. Deploying the entire model means the validation of the model ensures matching the average model outputs with real data use, showing the suitability of the proposed concept.

During our review of published literature regarding statistical modelling and forecasting of mobile money transactions in Kenya, we noticed several important gaps. As such, the necessity of this study is undoubted. Prior literature has investigated time series forecasting methods [8]. However, there are limited systematic studies based on Hybrid ARIMA-XGBOOST models designed to forecast the patterns and dynamics of Kenyan mobile money transactions. Despite presenting a promising sphere of capabilities, the hybridization of ARIMA and XGBOOST models is an area that needs to be investigated.

Research Methodology

Data Description

This study methodology and analysis used secondary data sourced from the Central Bank of Kenya (CBK), captured as cumulative monthly data from March 2007 to December 2023 in Kenya. This data repository presented an abundant resource for historical transaction records and gives insights into long-term trends and patterns of Mobile Money Transactions (CKB, 2024). The resulting beneficiaries of using secondary data from a highly respectable institution was saved in terms of costs and time,

great accessibility, and the ability to cover a judicious dataset. Nevertheless, validation and data quality measures were ensured to support the research conclusions. Complete source and citation of the Central Bank of Kenya was marked with the reliability of academic integrity and transparency in the entire research.

Data analysis for this research was undertaken using R studio (Version 4.4.0) programming language which was supplemented by multiple statistical packages and libraries to perform different analytical tasks. "forecast" was used for time series forecasting, "boost" for gradient boosting, "ggplot2" for data visualization, "tseries" for time series analysis, and "readxl" for reading Excel files are the main packages and libraries.

Hybrid Model Formulations

Auto-Regressive Integrated Moving Average (ARIMA)

The ARIMA (autoregressive integrated moving average) model is a model that has had a lot of developments from its very beginning days around the schools of time series issue. Its origins can be traced back to pioneering work by statisticians and researchers who laid the foundation for its constituent components: autoregressive (AR), moving average (MA), and Integrated, I. Autoregressive models were the first to be introduced by Yule (1927) and Walker (1931), describing how the current value of a time series could be seen in its motion from the past. In a similar manner with moving average models, which capture the short-run 'warts' of the data like fluctuations and random shocks, Box and Jenkins introduced the concept of moving average models in 1970. Integration, or differencing, was applied for handling the non-stationarity in the signal, which was property of the model that describes about change over time was while as whist the statistics characteristic is varied. The next big innovation that the methodology of autoregressive models made famous was the incorporation component nicknamed by Box and Jenkins; the tool turns nonstationary time series data into stationary time series [3]. The ARIMA model turned to be a unified framework, which was responsible for its authenticity and simplicity of employing in most situations. AR, MA and I components were combined within a single model gaining immense popularity with this approach of multiple sequences modeling.

The Auto-Regressive part was described according to equation 1;

AR_p

$$Y_t = c + \delta_1 Y_{t-1} + \delta_2 Y_{t-2} + \delta_3 Y_{t-3} + \cdots + \delta_p Y_{t-p} + \varepsilon_t \quad (1)$$

The Moving Average part of the time series model was described according to equation 2;

MA_q

$$Y_t = \mu + \vartheta_1 Y_{t-1} + \vartheta_2 Y_{t-2} + \vartheta_3 Y_{t-3} + \cdots + \vartheta_q Y_{t-q} + \epsilon_t \quad (2)$$

Therefore, the ARMA model was described according to equation 3 below;

$$Y_t = c + \delta_1 Y_{t-1} + \delta_2 Y_{t-2} + \delta_3 Y_{t-3} + \cdots + \delta_p Y_{t-p} + \varepsilon_t + \vartheta_1 Y_{t-1} + \vartheta_2 Y_{t-2} + \vartheta_3 Y_{t-3} + \cdots + \vartheta_q Y_{t-q} + \epsilon_t \quad (3)$$

When ARMA model undergoes treatments such as differentiation, integration, exponential smoothing and logarithmic expressions to reduce fluctuations in the data and improve the trends, and until stationarity is attained, the ARIMA(p,d,q) model was achieved in which p was the autoregressive order, q was the Moving Average order, while I was the integrated order and its value was given by the number of differentiations. This model was subjected to a stationarity test, used and the parameter scalar values of p, d and q was obtained from Box Jenkins test and auto.arima function and chosen as the best model.

Stationarity Test

The Augmented Dickey-Fuller (ADF) test is a statistical test that was used to determine whether a unit root was present in a time series. A unit root implies that a time series is non-stationary, that is its statistical features like mean and variance are time-varying. The ADF test also helps in checking whether the differencing of a time series is needed to achieve stationarity which is a precondition of modelling techniques like ARIMA. Regarding the Mobile Money Transaction data, the ADF test is applicable to check the stationarity of transaction volumes or values over time. If the p-value of the ADF test is less than a significance level (commonly 0.05), it will imply evidence against the presence of a unit root and supports that the data is stationary [3]. However, if the p-value will be greater than the

significance level the data is non-stationary therefore differencing may be needed to obtain stationarity before modelling. The ADF test mathematical equation representation will be;

$$\nabla y_t = \alpha + \beta t + \Upsilon y_{t-1} + \delta_1 \nabla y_{t-1} + \delta_2 \nabla y_{t-2} + \cdots + \delta_k \nabla y_{t-k} + \epsilon_t \quad (4)$$

Where;

∇y_t is the different time series at time t

α is a constant term.

β is a coefficient representing the trend component.

Υ is the coefficient on the lagged dependent variable.

δ_i are the coefficients on the lagged differences of the dependent variable.

ϵ_t is the error term.

Box Jenkins Test

The Box-Jenkins methodology, or the Box-Jenkins test, is a methodical approach involved in establishing the most suitable autoregressive integrated moving average (ARIMA) models for the forecasting of Mobile Money Transactions using historic data. The methodology involves three main stages: In this case, the main steps include: identification of the models to distinguish, determination of the parameters of the models, and model check or model diagnostic as highlighted. The specification involves determining the values of p for autoregressive, d for differencing, and q for moving average from plots of ACF (autoregressive) and PACF (partial autoregressive) graphs [3]. ML estimation is the process of determining the parameter values of the model that renders the given data most probable. These methods include maximum likelihood estimation which is widely used in contingent valuation studies. Diagnostic checking processes involve checking the residuals in order to determine how suitable the chosen model is if there is autocorrelation, non-normality or heteroskedasticity, which change is made through the use of other models or addition of more variables into the given model.

The Box-Jenkins test is aimed at identifying the re-current and non-seasonal mobile money transactions that will enable us to find the best fit of the MA, AR and integrated components of the Mobile Money Transactions such that the model selected will identify patterns in the Transactions data properly without over-fitting. In hauling out the mathematical formula for the Box-Jenkins test, the ARIMA model is avowed by its orders termed as p, d and q besides the parameters that are posited to be estimated.

ARIMA(p,d,q) (5)

Where:

p is the autoregressive order,

d is the differencing order,

q is the moving average order.

The Box-Jenkins test aims to determine the optimal values of

p, d, and q to construct the most suitable ARIMA model for forecasting Mobile Money Transactions.

Extreme Gradient Boosting Model (XGBOOST)

Extreme Gradient Boosting (XGBOOST), came about as a vital instrument in the world of machine learning, as it empowers one to deal with enormous amounts of data and also assures you of reliable predictions. XGBOOST, a highly optimized distributed gradients boosting library was produced by the minds of Chen Tianqi and Carlos Gestrin in 2011, with constant refinement by lots of scientists following that invention [5]. The unique concept of aggregating many states of low predictive accuracy models into one high accuracy model is incorporated into XGBOOST, and this allows a process of boosting that leads to the improvement of machine learning alone. Robust and effective, XGBOOST has recently arisen as a top-notch problem solver among researchers and domain experts in several areas of endeavor. The LightGBM advantage of factor significance evaluation of input elements so quickly, scalability, and the boosting of quantifiable algorithm highlights its significance in efficient optimization

data mining efforts. Especially with regard to the XGBOOST method in mobile payment analyses and predictions in Kenya, it is a very effective tool. Because of the widespread granularity and size related issues with transactional data, the fact that XGBOOST is a robust model with efficient handling of these makes it a logical choice for building forecast models. Through the iterative refinement and optimization of predictions, XGBOOST allows analysts and researchers to be able to derive operational and tactical insights and anticipate future developments when it comes to mobile money transactions [5]. This ultimately gives them the power to make educated decisions and develop strategies for the future of mobile money transactions.

The objective function of the XGBOOST algorithm will be subdivided into Loss function and regularization function given by;

$$Loss(y_i, \hat{y}_i) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (6)$$

While the regularization term $\Omega(h)$ is defined by

$$\Omega(h_m) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{k=1}^{T} \omega_k^2 \quad (7)$$

Where;

T is the number of leaves in the tree h.

ω_k are the output scores of the leaves.

γ controls the minimum loss reduction needed to slip an internal node.

λ is a regulation parameter.

The Loss function for XGBOOST is defined as follows:

$$L_{xgb} = \sum_{i=1}^{N} L(y_i, F(x_i)) + \sum_{m=1}^{M} \Omega(h_m) \quad (8)$$

Where;

$L(y_i, F(x_i))$ is the Loss function for the prediction

$\Omega(h_m)$ is a regularization term for the mth tree.

The regularization term $\Omega(h)$ is defined as;

$$\Omega(h_m) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{k=1}^{T} \omega_k^2 \quad (9)$$

Where;

T is the number of leaves in the tree h.

ω_k are the output scores of the leaves.

γ controls the minimum loss reduction needed to slip an internal node.

λ is a regulation parameter.

XGBOOST builds an additive approximation of the true function $F^*(x)$ as a weighted sum of functions.

$F_m(x)$ is model interaction m

P_m is the weight of the mth function

$h_m(x)$ is the prediction of the mth tree.

The training XGBOOST model has an objective function given by;

$$Objective = \sum_{i=1}^{n} Loss(y_i, \hat{y}_i) + \sum_{k=1}^{k} \Omega(f_k) \quad (10)$$

n is the number of observations.

$Loss(y_i, \hat{y}_i)$ measures the difference between actual and predicted residuals.

$\Omega(f_k)$ is the regularized term for the k[th] trees.

Prediction of Autoregressive Integrated Moving Average (ARIMA) residuals using Extreme Gradient Boosting can be estimated using the prediction equation as follows;

$$\hat{\eta}_t = \sum_{k=1}^{K} \hat{\eta}_t^{(k)} \quad (11)$$

$\hat{\eta}_t^{(k)}$ is the prediction of ARIMA residuals by the k[th] tree.

Building the Hybrid Model combining ARIMA and XGBOOST.

The Autoregressive Integrated Moving Average model is fitted to the time series data to obtain initial forecast. Then, the Extreme Gradient Boosting model is trained on the residuals of the Autoregressive Integrated Moving Average model to capture the additional patterns. Finally, the forecasted Autoregressive Integrated Moving Average residuals from the Extreme Gradient Boosting model is added to ARIMA forecasts to obtain the final hybrid model.

Fitted ARIMA

$$\hat{y}_t = c + \sum_{i=1}^{p} \delta_i y_{t-i} + \sum_{j=1}^{q} \vartheta_j \eta_{t-j} + \varepsilon_t \quad (12)$$

The fitted XGBOOST model for ARIMA residual is;

$$\hat{\eta}_t = \sum_{k=1}^{K} \hat{\eta}_t^{(k)} \quad (13)$$

Therefore, combing ARIMA and XGBOOST models for Forecasting Hybrid model that will give us the final forecasted values;

$$\hat{y}_H = \hat{y}_t + \hat{\eta}_t \quad (14)$$

The final hybrid model for forecasting mobile money transaction values was defined as follows;

$\hat{y}_H$ is the final forecasted value at time t from the hybrid of ARIMA and XGBOOST models.

$\hat{y}_t$ is the forecasted value from ARIMA model.

$\hat{\eta}_t$ is the forecasted value of the ARIMA residuals trained with the XGBOOST model.

Evaluation parameter of models

A model's real accuracy can be measured by comparing predicted and actual values. A variety of performance metrics can be performed to calculate accuracy. We used four prominent forecasting parameters to assess the predictive efficacy of our model: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Percentage Error (MPE), as follows:

MAE provides a measure of prediction accuracy by quantifying the average magnitude of errors in a set of predictions, without considering their direction this is described by the equations 16.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (16)$$

RMSE gauges how far the forecasts stray from the actual values by measuring the standard deviation of prediction errors and the equation is in 17.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (17)$$

MAPE indicates the average relative error size and gives the prediction accuracy as a percentage. See equation 18 below.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| X100\% \quad (18)$$

MPE indicates if the predictions often tend to be too high or too low by measuring the average bias in the predictions. See equation 19 below.

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{y}_i - y_i}{y_i}\right) X100\% \quad (19)$$

Where n is the number of observations and is the error between the forecasted and actual value. The mean of the actual sense of forecasting is arrived at by adding the individual differences between the forecast made and the actual forecasted value and dividing by the number of observations. RMSE is one of the most frequent measures applied when comparing the forecasted values by a model or an estimate to the observed ones, and it is described as the average of squared errors squared. MAPE quantifies the accuracy by percentage which is equal to the sum of simple percent difference form the actual values to the forecasted values divided by the sum of actual values throughout the period.

Analysis of Results and Discussions

Descriptive Statistics

The summary statistics provided reveal key insights into the distribution of Agent Cash Volume and Value. Both variables exhibit similar patterns, with means and medians close in value, indicating symmetric distributions. The minimum and maximum values, as well as the interquartile ranges (Q1 and Q3), suggest the presence of outliers, particularly evidenced by the notably higher maximum values relative to the upper quartiles. The standard deviations are relatively large compared to the means, indicating a spread of data points around the mean. These statistics collectively provide a snapshot of the central tendency, variability, and range of the data according to table 1 and table 2 below.

Measured in million Kenya shillings, Table 1 presents descriptive information for the total amount of cash in and cash out for agents. A minimum of 0.022, a maximum of 213.31, a standard deviation of 66.07312, a first quartile (Q1.25) of 35.465, a third quartile (Q3.75) of 153.1265, and a mean of 96.671 are among the important metrics. The standard deviation of 66.07312 million KES indicates that there is significant fluctuation in the average cash volume of 96.671 million KES, according to the figures. The large disparity between the lowest and maximum values, indicated by the median of 94.0595 million KES, points to a lean towards higher volumes.

Table 1: Descriptive statistics for Cash Volume.

Variable	Mean	Median	Minimum	Maximum	Standard Deviation	Q1.25	Q3.75
Volume	96.671	94.0595	0.022	213.31	66.07312	35.465	153.1265

Descriptive statistics for the total agent cash in and cash out, expressed in billions of Kenya Shillings, are given in Table 2. With a standard deviation of 210.8436 billion KES and a mean value of 269.6075 billion KES, there is a considerable degree of variability. With values ranging from 0.064 to 788.35 billion KES, the median is 237.618 billion KES. The first and third quartiles are 93.076 and 366.2665 billion KES, respectively. The standard deviation of 210.8436 billion KES illustrates the significant diversity present in the statistical statistics, which indicate an average agent cash value of 269.6075 billion KES. With cash values varying significantly from 0.064 to 788.35 billion KES, the distribution appears to be slightly skewed, as indicated by the median value of 237.618 billion KES. The middle 50% of the data is dispersed over a wide range, as seen by the quartile values (Q1.25 at 93.076 and Q3.75 at 366.2665 billion KES), which indicate a significant variation in agent cash values.

Table 2: Descriptive statistics for Cash value.

Variable	Mean	Median	Minimum	Maximum	Standard Deviation	Q1.25.	Q3.75.
Value	269.6075	237.618	0.064	788.35	210.8436	93.076	366.2665

Time Series Plots on Total Agent Volume

The plot in Figure 1 below indicates that the data shows a consistent positive trend across time. The volume of total agent cash in cash out in volume in Million Kenya Shillings have consistently increased since the inception of the Mpesa. The curve appeared sharp between the December, 2019 and April

2020, and this was attributed by the outbreak of Corona Virus Disease, which made the government of Kenya encourage the use of cashless transactions.

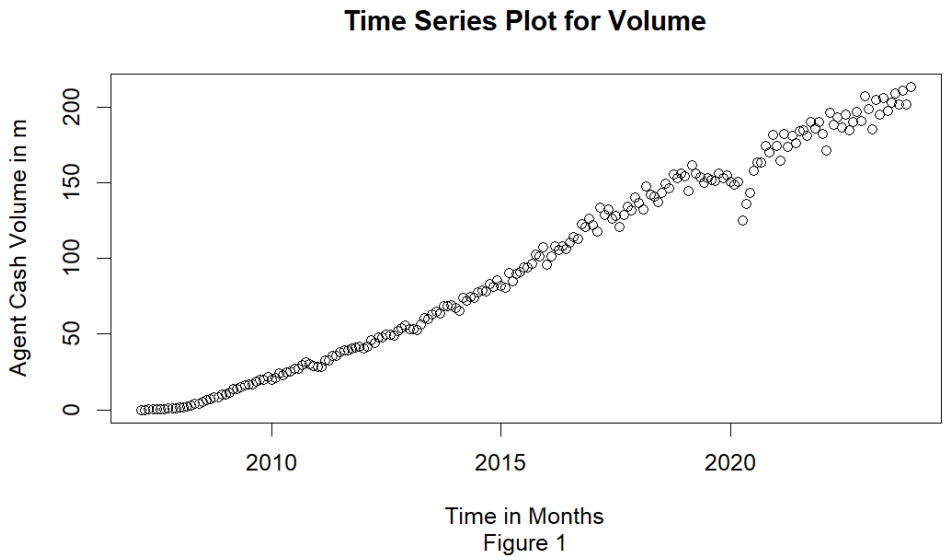

Figure 1: Time Series plot for Total Agent Cash Volumes.

Time Series Plots on Total Agent Value

The plot in Figure 2 below was nearly similar and indicated that the data shows a consistent positive trend across time. The value of total agent cash in cash out in value in Billion Kenya Shillings have consistently increased since the inception of the Mpesa. The curve appeared sharp between the December, 2019 and April 2020, and this was attributed by the outbreak of Corona Virus Disease, which made the government of Kenya encourage the use of cashless transactions.

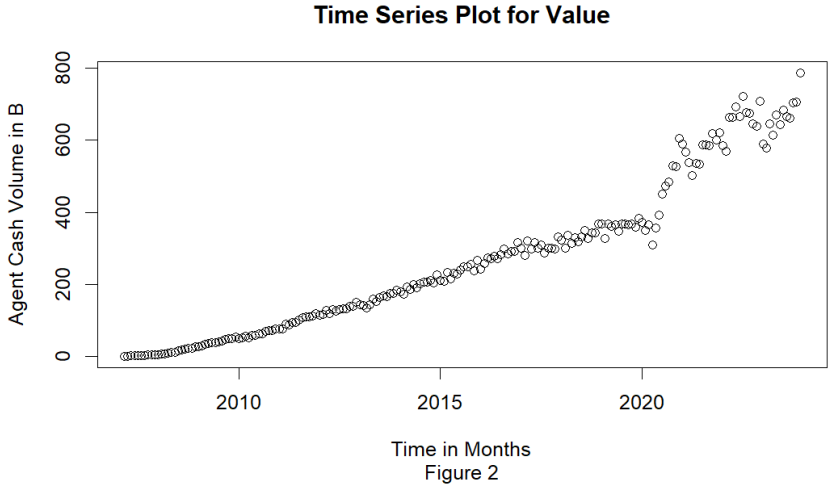

Figure 2: Time Series plot for Total Agent Cash Values.

Box Plots for Checking Variability of the Dataset

Regarding the boxplots for Agent Cash Volume and Value, they would likely illustrate visually the spre ad and distribution of the data, offering additional insights beyond the summary statistics alone. The b oxplot in Figure 3 would likely reveal the presence of outliers towards the upper end of the distribution, given its larger maximum value relative to the upper quartile. Similarly, the boxplot in Figure 4 would li kely exhibit a similar pattern. Comparing the boxplots for both variables would offer a visual compariso n of their distributions, aiding in identifying similarities and differences between the two. The presence of outliers can be exhibited to the outbreak of Corona Virus disease between late 2019 to the end of 2 022 that reduced to the use of cash money to Mobile money transaction paving way to a new lifestyle.

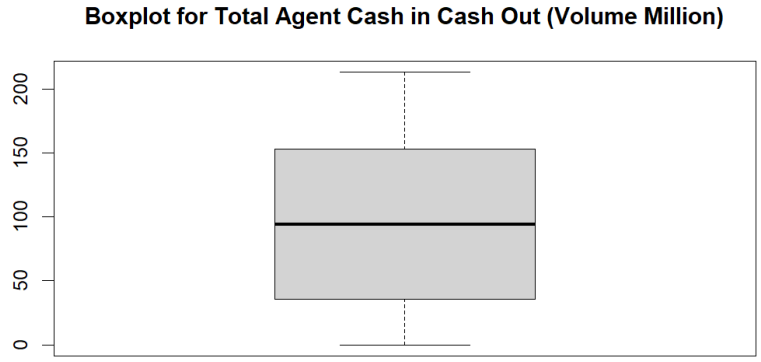

Figure 3

Figure 3: Boxplot for Agent Cash volumes.

Boxplot for Agent Cash volumes.

It is showing some form of outliers as can be seen from the dot point above the whisker.

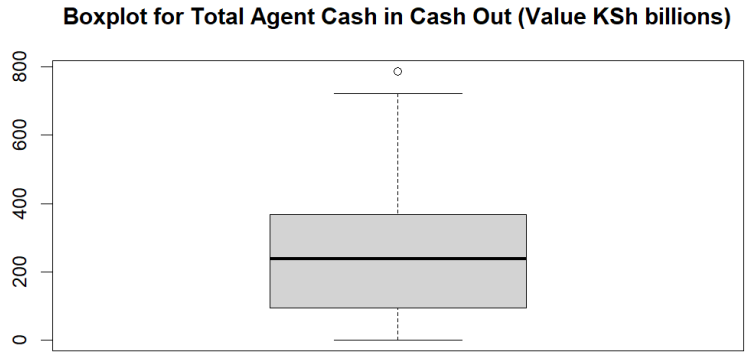

Figure 4

Figure 4: Boxplot for Agent Cash values

Hybrid Model Formulation

The results of the Augmented Dickey-Fuller (ADF) test on the differenced data in tables 3 and 4 indica ted that the data is stationary. With a highly significant test statistic and a p-value of 0.01, below the ty pical significance level of 0.05, we reject the null hypothesis of non-stationarity. This suggests that diff erencing has successfully removed trends and seasonality from the data, making it suitable for further analysis and forecasting.

Table 3: ADF test for Cash Volume.

Test	Test Statistic	p-value	Lag Order	Alternative Hypothesis
Augmented Dickey-Fuller Test	-7.9038	0.01	5	Stationary

The test statistic value for total agent cash in cash out value is -6.9807 compared to -7.9038 or the tot al agent cash in cash out volume. However, in both cases the p-values are less than the significance l evel of 5%, making the test to be statistically significance.

Table 4: ADF test for Cash Value.

Test	Test Statistic	p-value	Lag Order	Alternative Hypothesis

Augmented Dickey-Fuller Test	-6.9807	0.01	5	Stationary

ARIMA Model Components Identification

The ARIMA components for both table 5 and table 6 provided insights into the time series modelling structures. For table 5, the ARIMA model is specified as (1,1,0) (1,0,0) [12], indicating first-order differencing, autoregressive seasonal component with a lag of 12, and a drift term. Interestingly, the coefficients indicate a negative first order autoregressive term and positive seasonal auto-regressive term, though the drift is negative. In the same manner for table 6, the selected ARIMA model is somewhat more complex as (3,1,1) (2,0,0) [12]. There are AR and MA of multiple orders and seasonal components in this model. Co efficiencies give information about strength and direction of the link between the lagged values of a variable. The next is the AIC and BIC that measure the fitness level and models with the smallest figures in the index are usually considered to be the best.

Table 5: ARIMA components for Cash Volume

Model	Coefficients	Std. Errors	Sigma^2	Log Likelihood	AIC	AIC	BIC
ARIMA (0,1,1) (1,0,0) [12] with drift	-0.3340, 0.5668, 0.9689	0.0774, 0.0666, 0.4478	19.89	-586.6	1181.2	1181.4	1194.41

These ARIMA components help in establishing the basic structure and characteristics of the data in order to provide a good fit for forecasting and making decisions in mobile money transactions for Kenya. While the BIC values are more useful while choosing the right model, the AIC was appropriate for employing while in search of the right model that would enhance the prediction of the future observations (Chakrabarti & Ghosh, 2011).

Table 6: ARIMA components for Cash Value

Model	Coefficients	Std. Errors	Sigma^2	Log Likelihood	AIC	AIC	BIC
ARIMA (0,1,0) (1,0,0) [12]	0.6358	0.0627	365.4	-880.87	1765.74	1765.8	1772.35

Autocorrelation Function and Partial Autocorrelation Function

The ACF and PACF plots help in identifying the level of dependency in the differenced time series data of Agent Cash in Cash Out as the Volume. In the ACF plot there is evidence of first order autocor-relation and a possible seasonal variation at lag 12. The PACF plot enables for determining the number of an autoregressive process, since we observe a significant bar at the lag equal to 1, and bars at the lag equal to 2 and 12 that are also rather high. The above plots justify non-seasonal and seasonal autocorrelations in the differenced series which corroborates with the ARIMA model estimated above. Summarily, these plots enable one determine the right parameters for the ARIMA model and helps in understanding the temporal patterns depicted in the Figure 5 below.

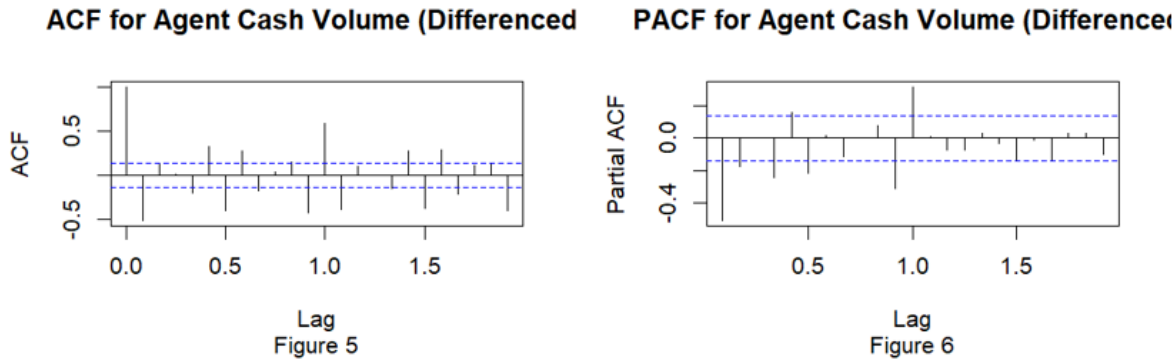

Figure 5: ACF and PACF plot for Agent cash Volumes.

Box-Jenkins test for Total Agent Volume

The Box-Jenkins test is conducted to assess the adequacy of the ARIMA model by examining the auto correlation of the residuals. In this result, the Box-Jenkins test statistic, X-squared, is large (380.31) an d the p-value is extremely small (p-value < 2.2e-16), indicating significant evidence against the null hy pothesis of no autocorrelation. This result suggests that there is still autocorrelation present in the resi duals of the ARIMA model in table 7 after differencing, indicating that the model adequately capture all the temporal patterns in the data.

Table 7: Box Ljung test for Cash Volume

Test	Test Statistic (X-squared)	p-value	Degrees of Freedom	Data
Box-Ljung Test	380.31	< 2.2e-16	20	diff_ts_volume

ACF and PACF for Total Agent Cash Value

The provided R code generates plots for the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the differenced time series data in figures below. These plots are essential for ident ifying potential ARIMA model parameters. The ACF plot shows the correlation between the series and i ts lagged values, while the PACF plot displays the correlation between the series and its lagged value s after removing the correlations explained by the intermediate lags. By examining these plots, we can identify the lag orders for the AR and MA components of the ARIMA model. If there are significant corr elations at the initial lags in the ACF plot followed by a sharp cutoff in the PACF plot, it suggests an AR component, while the opposite pattern indicates an MA component. These plots provide valuable insig hts into the temporal patterns and dependencies present in the Mobile Money Transaction data, aiding in the selection of appropriate ARIMA model parameters.

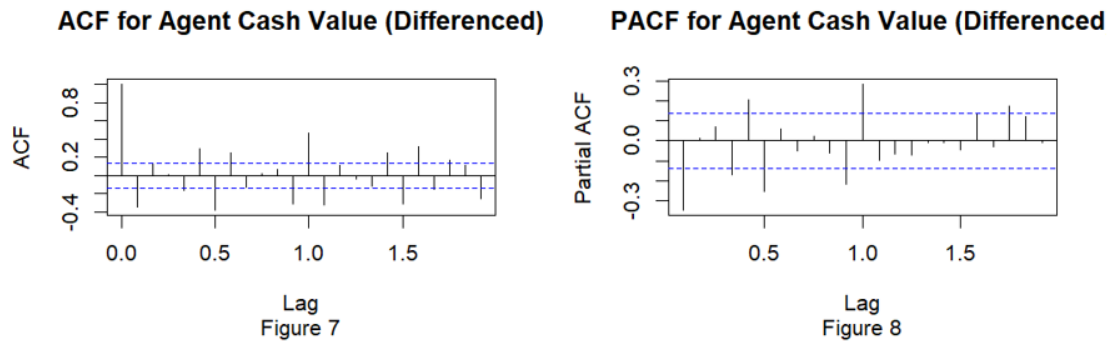

Figure 6: ACF and PACF plot for Agent cash Values.

Box-Jenkins Test for Total Agent Cash Value

The provided Box-Jenkins test for the differenced in table 8 time series data assesses the presence of autocorrelation at various lag intervals. The test statistic, X-squared, is compared to a chi-squared dist ribution with degrees of freedom equal to the number of lags specified. In this case, the test yields a v ery low p-value (< 2.2e-16), indicating strong evidence against the null hypothesis of no autocorrelatio n. Therefore, we reject the null hypothesis and conclude that there is significant autocorrelation presen t in the differenced Agent Cash in Cash out in Value of the transaction data at the specified lag interval s. This finding suggests that an ARIMA model may be appropriate for modelling this time series data, a s it accounts for autocorrelation and can capture the underlying temporal dependencies.

Table 8: Box Ljung test for Cash Value

Test	Test Statistic (X-squared)	p-value	Degrees of Freedom	Data
Box-Ljung Test	260.24	< 2.2e-16	20	diff_ts_value

Training the Model

In the following segment of the R code, the Head of Vodafone Mobile Money Traders is trained on XG BOOST models with various forms of loss functions to forecast both the total volume of Agent and valu e in billions of mobile money transactions simultaneously. The models were successfully trained with t he evaluation metric stated as root mean square error or RMSE. Each model was trained with a differe nt loss function, including reg: Line, regression model: logistic regression, and regression model: twee die. These created models were printed to realize and show their characteristics. Through this analysi s, it becomes clear that XGBOOST supports all kinds of loss functions for the purpose of regression th at provide less restriction on the model optimization and its corresponding evaluation. It might be requi red to talk further about the usage of the parameters in order to get a better picture because different l oss functions were not as effective on the test data set.

Model Evaluation

The model evaluation parameters provide valuable insights into the performance of the predictive mod el. The MAPE (Mean Absolute Percentage Error) value of 0.362 indicates that, on average, the model' s predictions deviate from the actual values by approximately 0.362%. Given that the standard deviati on for Total Agent Cash Volume is 66.07312 and for Total Agent Cash Value is 210.8436, this level of e rror seems quite low and suggests that the model is making accurate predictions with respect to the va riability in the target variables. The standard deviation of 0 indicates that the values of the predicted an d actual sales are very close and the RMSE (Root Mean Squared Error) value of 0 also approves the s ame idea. 2019 supports this, and shows that on average the error of prediction is small relative to the spread of the data. As for the quantitative evaluation, the MPE (mean percentage error) and MAE (me an absolute error) have been reported with the value of 0. 362 and 0. The metrics of the calibration set that support the effectiveness of the model include the number of correctly identified user images, 918 53, and minimal and absolute errors of 0.0065 and 0.0402, respectively. In general, these evaluation m etrics show that the model gives quite accurate results in estimating Total Agent Cash Volume and Tot al Agent Cash value in Term IR measurements and shows variance with respect to the variability that i s present in the data.

Table 9: Model Evaluation parameters.

Model Evaluation Parameter	Evaluation value (%)
MAPE	0.362
RMSE	0.2019
MPE	0.362
MAE	0.919

Discussions of the Analysis Results

The specific aim of the current study was to model the Mobile Money Transactions in Kenya using the hybrid ARIMA-XGBOOST technique. This objective was achieved through creating a Hybrid ARIMA-X GBOOST model that describes the transaction parameters and behaviours. The first step in analysis i nvolved checking on the stationarity of data through the ADF test and the results for the stationarity of the data were established that Data differencing has led to elimination of trends and seasonality for fur ther analysis and Data forecasting. The components of the ARIMA model were then determined for “A gent Cash Volume” and “Agent Cash Value,” which explained the time series modelling architecture’s c haracteristics [11]. These components provided the basis for identifying the critical forces within the giv en data set, as well as the patterns that shaped its development; thus, it was possible to make approp riate forecasts and estimating risks.

ACF and PACF plots were very useful in determining the value of parameters of the ARIMA model and perfect confirmation of non-seasonal and seasonal autocorrelation in the data. Acute test, such as Box -Jenkins test, further investigated the suitability of ARIMA model by checking the residual autocorrelati on which indicated that all temporal features and characteristics of the data were endogenously model led by the ARIMA model. Detailed experiments were performed with different combinations of loss fun ctions when training XGBOOST models; this proved that the model was very flexible, and easily adopt ed different optimization strategies to provide good predictions of both “Agent Cash Volume” and “Age nt Cash Value” [7]. The components plot helped in the analysis of individual components of the time se ries in order to understand general trends, seasonal patterns, and fluctuation all of which were importa nt in the decision-making regarding forecasting of mobile money transactions.

Conclusion and Recommendations

Conclusions.

Overall, this research achieved the aim of creating and evaluating a blend of ARIMA-XGBOOST in des cribing the trends and changes in Kenyan mobile money transactions. The main goals were to develop a HA-MDL based on the advantages of both ARIMA and XGBOOST, as well as to explore temporal pa tterns of historical transaction records to predict their future values. Thus, along with offering a detailed picture of the mobile money market in Kenya, which was the first objective of this study, this paper acc omplished the second objective through thorough statistical modelling and forecasting.

The first formulated objective that was aimed at creating a hybrid ARIMA-XGBOOST model was acco mplished by utilizing the ARIMA model to identify linear moving and seasons within the “Total Agent Ca sh Volume” and “Total Agent Cash Value”. The values for the parameters of the ARIMA models were d etermined based on the characteristics of the ACF and PACF plots, while reliability of the adopted mod els was checked using the Box-Jenkins test. Therefore, the XGBOOST model was used in order to fit on the residuals of the fitted ARIMA model in order to account for non-linearity. The proposed hybrid m ethodology helped in enhancing the forecasting accuracy, which proves the hypothesis that the integra tion of ARIMA for linear and non- linear residuals as well as XGBOOST with its strong capability of dea ling with non-linear trend.

Recommendations

1. To further enhance the practicality of the model, the analysis will be continually updated with new data, thereby verifying its predictive ability and applicability in delivering actionable information for stakeholders’ strategic management and policy-making processes.
2. In the same regard, to increase the effectiveness of the Hybrid ARIMA-XGBOOST model in identifying patterns and dynamics in mobile money transactions, more investment in data collection and integration should be enhanced across the financial institutions and the regulatory authorities such as the Central Bank of Kenya, Kenya Revenue Authority. This should include but not be limited to; Actual historical transaction records and/or real time data feed from diverse cross-border mobile money platforms.

References

[1] Kirui, R. K., & Onyuma, S. O. (2015). Role of mobile money transactions on revenue of microbusiness in Kenya. *European Journal of Business and Management*, *7*(36), 63-67.

[2] Ndirangu, L., & Nyamongo, E. M. (2015). Financial innovations and their implications for monetary policy in Kenya. *Journal of African Economies*, *24*(suppl_1), i46-i71.

[3] Shumway, R. H., Stoffer, D. S., Shumway, R. H., & Stoffer, D. S. (2017). ARIMA models. *Time series analysis and its applications: with R examples*, 75-163.

[4] Kullaya Swamy, A., & Sarojamma, B. (2020). Bank transaction data modelling by optimized hybrid machine learning merged with ARIMA. *Journal of Management Analytics*, *7*(4), 624-648.

[5] Fofanah, A. J. (2021). Machine learning model approaches for price prediction in the coffee market using linear regression, XGB, and LSTM techniques—*International Journal of Scientific Research in Science and Technology*, (6).

[6] Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, *37*(1), 388-427.

[7] Islam, S. F. N., Sholahuddin, A., & Abdullah, A. S. (2021). Extreme gradient boosting (XGBoost) method in making forecasting application and analysis of USD exchange rates against rupiah. In *Journal of Physics: Conference Series* (Vol. 1722, No. 1, p. 012016). IOP Publishing.

[8] Mwila, R., & Kunda, D. (2022). Using data mining techniques, a predictive model to determine the growth of mobile money transactions in Zambia. *International Journal of Data Mining and Bioinformatics*, *27*(1-3), 139-170.

[9] Fabregas, R., & Yokossi, T. (2022). Mobile money and economic activity: Evidence from Kenya. *The World Bank Economic Review*, *36*(3), 734-756.

[10] Suri, T., Aker, J., Batista, C., Callen, M., Ghani, T., Jack, W., ... & Sukhtankar, S. (2023). Mobile money. *VoxDevLit*, *2*(2), 3.

[11] Zhan, C., Liu, Y., Wu, Z., Zhao, M., & Chow, T. W. (2023). A hybrid machine learning framework for forecasting house prices. *Expert Systems with Applications*, *233*, 120981.

Definitions of Terms

Total Agent Cash in Cash Out (Volume Million): The volume of cash transactions in millions performed by mobile money agents. Reflects the magnitude of cash transactions occurring through mobile money agents. This variable helps understand the demand for cash-related services, providing insights into the cash economy's reliance on mobile money.

Total Agent Cash in Cash Out (Value KSh billions): The total value of cash transactions in Kenyan Shillings (KSh) billions performed by mobile money agents. Indicates the economic value of mobile money transactions. Higher values suggest a significant economic contribution, influencing liquidity and financial inclusion.