AN EFFICIENT TIME SERIES MODEL FOR TAX REVENUE FORECASTING: A CASE OF NIGERIA

A dissertation presented to the department of Computer Science,
African University of Science and Technology, Abuja-Nigeria
In partial fulfilment of the requirements for a Master's degree in
Management of Information Technology

By

Ajisola Ayotola Segun (41024)

Abuja, Nigeria

MAY, 2023

CERTIFICATION

This is to certify that the thesis titled AN EFFICIENT TIME SERIES MODEL FOR TAX REVENUE FORECASTING: A CASE OF NIGERIA submitted to the school of postgraduate studies, African University of Science and Technology (AUST), Abuja, Nigeria for the award of the Master’s degree is a record of original research carried out by Ajisola Ayotola Segun in the Department of Computer Science.

12.05.2023

SIGNATURE PAGE

AN EFFICIENT TIME SERIES MODEL FOR TAX REVENUE FORECASTING: A CASE OF NIGERIA

By

Ajisola Ayotola Segun

A THESIS APPROVED BY THE COMPUTER SCIENCE DEPARTMENT

RECOMMENDED:

Supervisor: Dr. Rajesh Prasad

Head of Department: Dr. Rajesh Prasad

APPROVED: ______________________________

Chief Academic Officer

____

Date

ABSTRACT

Federal government independent revenue, non-oil revenue and oil revenue are some of the different sources of money for the Nigerian government. The sources of tax revenue include Pay as You Earn (PAYE), Stamp Duty (STD), Companies Income Tax (CIT), Value Added Tax (VAT), Personal Income Tax, and Petroleum Profit Tax (PPT). Due to the fact that taxes are now one of Nigeria's main sources of income, it is crucial to understand what to expect in terms of their amount. This will either help identify how to improve the country's budget or how to align it with the country's economic situation, depending on the current economic climate. This study makes use of monthly data collected over a period of time based on Federal inland revenue service (FIRS) source data related to earlier collections from a variety of tax categories between the years 2010 and 2021. To determine the optimal model, this study analyzed the projected values and model accuracy from three models multivariate Linear Regression (MLR), seasonal autoregressive integrated moving average (SARIMA) and multi-variate long short-term memory networks (LSTM). Because we could predict using multiple independent variables, both LSTM and MLR fared better. The LSTM model had a R^2 score of 98.9% and an adjusted R^2 score of 98.8%. Our findings indicate that multi-variate long short-term memory networks can be used to forecast tax revenue with reasonable accuracy and the multivariate Linear Regression comes close when multiple independent variables are used. This can further be enhanced by using other macro-economic factors for greater accuracy.

Keywords: Tax revenue; Time series model; SARIMA model; LSTM model; MLR model; RMSE test; Granger causality; R2 score; MAPE; MSE; MAE.

ACKNOWLEDGEMENTS

I wish to express my deep gratitude to Prof. Prasad for his constant guidance, motivation, and support during the course of my research. His broad expertise, talent, and insights have had a significant impact on my academic work and professional growth. I appreciate his advice and the ongoing inspiration, challenges, and motivation he has given me.

I also want to express my gratitude to my departmental colleagues for their unwavering advice, recommendations, and words of support. I am grateful for the opportunities they have provided for me to advance professionally and provide something to the scientific community.

My sincere gratitude also extends to my friends and family for their unwavering love, support, and tolerance during my academic endeavors. Their support has been crucial in enabling me to get through the challenges and to fully appreciate the experience

Finally, I express my gratitude to God for his blessings, direction, and grace in enabling this accomplishment.

DEDICATION PAGE

I write my dissertation as a dedication to my family and best friend’s. Special thanks go out to my devoted parents and best friend, whose words of support and push for persistence continue to reverberate in my ears. I also dedicate this dissertation to all of my close friends, close colleagues at work and my adopted family for their help and encouragement during the writing process.

LIST OF ABBREVIATIONS AND TERMS

FIRS	Federal inland revenue service
CIT	Companies Income Tax
VAT	Value Added Tax
STD	Stamp Duty
PPT	Petroleum Profit Tax
PIT	Personal Income Tax
AI	Artificial Intelligence
SARIMA	Seasonal autoregressive integrated moving average
LSTM	Long Short-Term Memory
MLR	Multivariate Linear Regression
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
RMSE	Root Mean Squared Error
MSE	Mean Squared Error

Table of Contents

List of Figures

List of Tables

CHAPTER 1 – INTRODUCTION

1.1. Introduction

To supply citizens with public goods and services, governments need financial resources. Governments must first decide whether there are sufficient financial resources before deciding how much of these goods should be provided. Federal government independent revenue, non-oil revenue and oil revenue are the three categories into which Nigerian government revenue is divided. Pay as You Earn (PAYE), Stamp Duty (STD), Companies Income Tax (CIT), Value Added Tax (VAT), Personal Income Tax, and Petroleum Profit Tax (PPT) and other taxes generate revenue (FIRS, 2022). According to statistics, the third quarter has the biggest revenue collection in history. Taxation has become one of Nigeria's most important sources of revenue, thus knowing how much to expect in the future will help either improve or align the country's budget. The fundamental purpose of this research is to forecast Nigeria's future income tax revenues. It is critical for a country to understand the accuracy of its expected income tax revenue to maximize the national budget. Tax revenues are the fundamental statistics required in development planning and budget preparation. Currently we have several models that has been used for forecasting and prediction both classical models like Simple Exponential Smoothing, Seasonal autoregressive integrated moving average, autoregressive integrated moving average, vector autoregressive etc. and machine learning models like decision trees, Recurrent Neural Network, Multilayer Perceptron's (MLP), and Long Short-Term Memory (LSTM) network models. Predicting tax revenue is a challenging task, and even more so when multiple variables are considered (Jang, 2019).

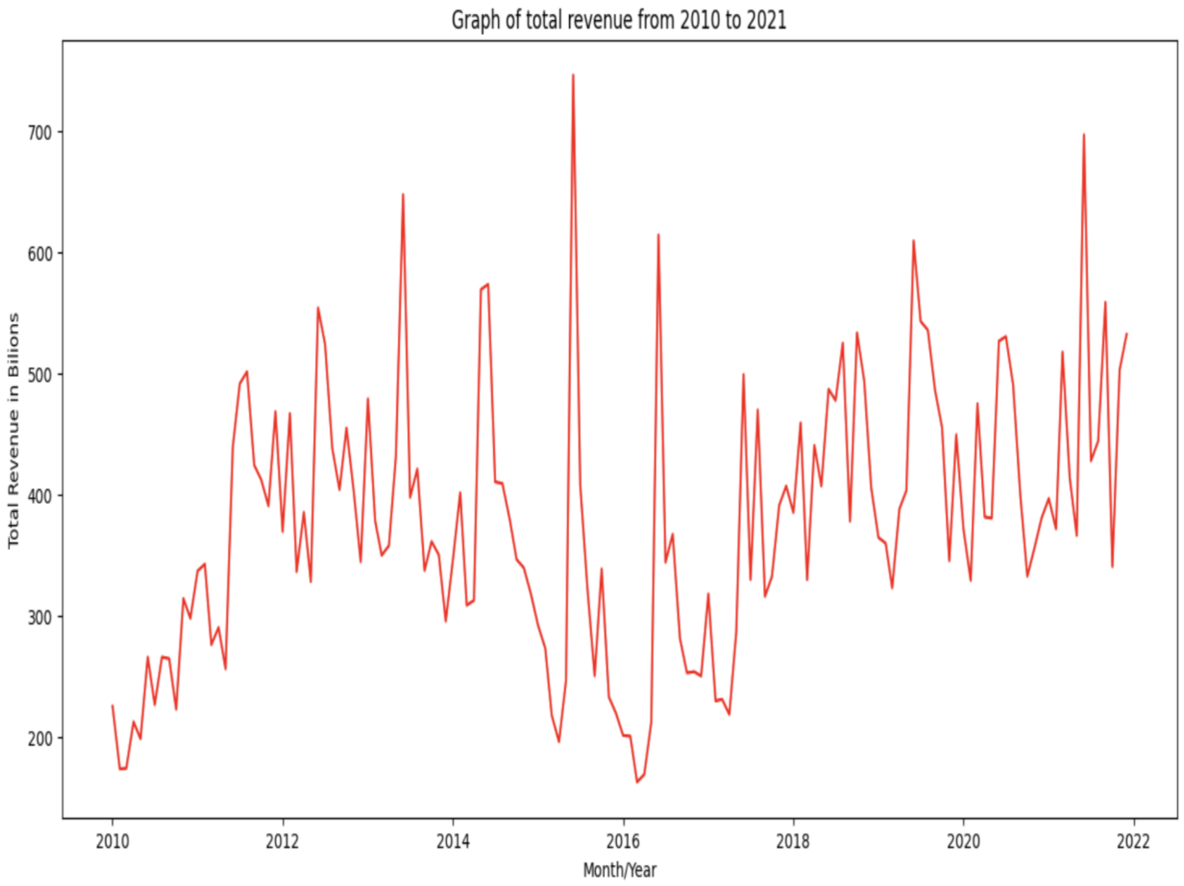


Figure 1. Total tax revenue - 2010 – 2021

Due to the country's reliance on oil money and the volatility of the oil market, tax revenue forecasting in Nigeria is a significant concern. Additionally, Nigeria has a relatively low tax-to-GDP ratio compared to other countries, which means that there is room for improvement in terms of tax collection. One of the key challenges in predicting tax revenue for Nigeria is the lack of reliable data. According to a study by the International Monetary Fund, Nigeria has a relatively weak statistical capacity, and the availability of data on government revenue and expenditure is limited. This makes it difficult to build accurate models for tax revenue prediction. Predicting tax income in Nigeria has been the subject of several studies, but most of them have focused on a single variable, such as oil revenue or Value Added Tax (VAT). For example, a study by (Ajayi et al., 2013) predicts Nigerian oil revenue using time-series analysis and found that the ARIMA model performed the best.

1.2 Research Problem

For the Nigerian government to minimize underfunding and overfunding, accurate income forecasting is crucial. Forecasting makes use of the data that is currently accessible in addition to other analytical approaches in order to forecast a variable's future value. Here, as tax revenues are the subject, we are interested in discovering which model is the most successful overall and will be applied to forecast future revenue by comprehending the variables that affect the generation of taxes. Pay-As-You-Earn (PAYE), Petroleum Profit Tax (PPT), Companies Income Tax (CIT), Personal Income Tax (PIT), Value Added Tax (VAT), Stamp Duty (STD), etc. are just a few of the different sorts of taxes we have. Which of them has a major impact on the generation of tax revenue is what we want to discover. We will be using multivariate Linear Regression (MLR), multi-variate long short-term memory networks (LSTM) and seasonal autoregressive integrated moving average (SARIMA).

1.3 Research Questions

1. Can total revenue be predicted with just past total revenue data?
2. Which of the three chosen models (SARIMA, LSTM, MLR) will perform better?
3. Is the discrepancy between the actual and predicted values significant?
4. What is the optimal SARIMA model for Total Tax Revenue?
5. The following variables: Is Granger Causality Present?
    a) Petroleum Profit Tax
    b) Value Added Tax
    c) Capital Gain Tax
    d) Companies Income Tax

1.4 Objective of the study

This study's primary objective is to introduce and contrast the use of multi-variate long short-term memory networks (LSTM), seasonal autoregressive integrated moving average (SARIMA) and multivariate linear regression (MLR) models in forecasting and predicting the overall yearly tax revenue of previous years and to compare actual and predicted values using Python's data science libraries. In order to track Nigeria's total tax revenue, the purpose of this research is to make projections and zero in on the independent variables that have the potential to accurately predict the dependent variable. It seeks to create a model with less errors than the others made along the process. The government may use the findings of this study because it controls the national budget and has the authority to decide how to spend the taxes it has collected to boost the economy of the nation.

1.5 Research Gap

While several studies have explored the effectiveness of different forecasting models for tax revenues in various countries, there is still a need for further research to develop accurate and reliable tax forecasting models specifically for Nigeria. The following research gaps should be addressed in a thesis on Nigerian tax forecasting model:

1. Lack of consideration of political and social factors: Political and social factors can significantly impact tax revenues in Nigeria. For instance, changes in government policies, social unrest, and other factors could lead to fluctuations in tax revenues. A forecasting model that does not account for these factors may not be able to provide accurate forecasts.
2. Limited research on new modeling techniques: There may be a lack of research on new modeling techniques that could be applied to tax forecasting in Nigeria. For instance, machine learning and artificial intelligence techniques could be leveraged to improve the accuracy of tax revenue forecasts, but there may be limited research on how to apply these techniques in the Nigerian context.
3. Limited availability of data: One of the key challenges of building robust tax forecasting models is the availability and quality of data. In Nigeria, there may be a lack of reliable historical data on tax revenues, which could make it difficult to develop accurate forecasting models.

In summary, the absence of trustworthy data, the economy's volatility, and the informal nature of the economy make it difficult to forecast tax collection in Nigeria. It is also challenging to locate earlier study on the topic because there is a gap in the literature.

1.6 Contribution

The importance of this research is multifaceted; for instance, government income forecasting studies for Nigeria are still in their infancy. This study aims to fill that vacuum in the literature, which will be helpful for future research studies. When predicting Nigeria's total revenue, we took into account seasonal characteristics; this is an appropriate and useful forecasting technique (Rob J Hyndman & George Athanasopoulos, n.d.). We have looked at total revenue projections and the effects of Value Added Tax (VAT), Companies Income Tax (CIT), and Petroleum Profit Tax (PPT) on total tax revenues in Nigeria. The findings of this study serve as a foundation for creating policies that are consistent with the interests of the general public, private institutions, and decision-makers. Another critical element is the type of data utilized in this analysis, which covered the period from January 2010 to December 2021 and comprised 144 records for various tax categories. Nigeria typically retains quarterly data, although the country is making progress in this area.

1.7 Organization of the thesis

i. **Chapter 1:** Introduction; focuses on shining light on the types of taxes in Nigeria and why they are crucial to economic progress. In addition, this chapter describes the aims and objectives of the research, the research questions, and the academic and practical significance of this study.

ii. **Chapter 2:** Literature Review; This chapter offers a detailed review of the available literature on tax revenue forecasting and modeling as it pertains to this study. It also shows the created theoretical framework of the research, which provides the foundation for the overall structure of the research by elaborating on the numerous theoretical facets of the subject at hand.

iii. **Chapter 3:** Proposed methodology; In this chapter, we delve deep into the theoretical underpinnings of the SARIMA, LSTM, and MLR models. In addition to this, we talk about the correctness of the model and how it was measured.

iv. **Chapter 4:** Experimental results stating the dataset; This addresses the application of the methodologies described in chapter 3 on the total income of Nigeria based on monthly revenue from 2010 to 2021 and monthly individual tax types.

v. **Chapter 5**: Discussion and Conclusion. Finally, findings and suggestions will be offered, along with observations and suggestions for how the research may be improved.

CHAPTER 2 – LITERATURE REVIEW AND TERMINOLOGIES

2.1. Introduction

The degree to which individuals comprehend the relevant data or variable has an impact on the accuracy of future forecasts. Predicting future events will always be limited by the predictive power of the approach or model used. Given the expansion in data availability brought on by improved data collection efficiency and technology, quantitative methods are currently widely used. Some causes are already known to limit predictive capacity, but there are unknown occurrences that could possibly introduce uncontrollable errors to future estimates (Siami-Namini et al., 2018). Tax revenue prediction or forecasting involves using various methods and techniques to estimate the future revenue that a government will collect from taxes. The goal is to provide policymakers and budget analysts with an accurate and reliable estimate of the amount of money that will be available for public spending, so that they can make informed decisions.

One approach that has been widely used in the literature is the use of time series models. These models can use techniques like moving averages, exponential smoothing, and ARIMA models to forecast future values of a variable based on its past values. For instance, a study by (Chen & Chen, 2017) the authors used an ARIMA model to predict tax revenue in Taiwan. They found that the model performed well, and that it was able to accurately predict future values of tax revenue.

Another approach that has been used in the literature is the use of econometric models. Using these models, it is possible to estimate the relationship between tax revenue and different economic metrics like GDP and inflation. For instance, a study by (Martinez-Vazquez & McNab, 2010), the authors used panel data regression to estimate the relationship between tax revenue and GDP in Latin American countries. They found that GDP was positively related to tax revenue, and that this relationship was robust across different countries and time periods.

A third approach that has been used is the use of Machine Learning and Artificial Intelligence models, such as Neural networks, Random Forest, etc. These methodologies provide more flexibility and can be applied on big datasets. For example, in a study by (Yaser S. Abu-Mostafa et al., 2012) used a Neural Network to predict the US tax revenue, which are found to be more accurate than traditional time series models.

Overall, the literature on tax revenue prediction and forecasting is diverse, with many different methods and approaches having been used. It is significant to remember that no approach is flawless, and the selection of a method will be influenced by the particular situation and the data at hand. To choose the best research methodology, it is crucial to speak with professionals in the fields of economics and public finance. (Matta et al., 2021).

2.2. Review studies on SARIMA models

Tax revenue forecasting frequently use the time series forecasting technique known as SARIMA (Seasonal Autoregressive Integrated Moving Average). This approach models the trend and seasonality of the time series data. The seasonal element is a part of an extension of the ARIMA model. SARIMA models have been applied in numerous research to anticipate tax income. One benefit of using SARIMA models for estimating tax revenue is their ability to detect subtle patterns in the data, such as seasonality and trend.

(Otu et al., 2014) applied the Seasonal ARIMA model in order to make projections on Nigeria's inflation rates. For the purposes of this investigation, data were gathered between the months of November 2003 and October 2013 (One hundred and twenty observations). The primary objective was to acquire a model that accurately portrayed the information, and the secondary objective was to create estimates regarding the rate of inflation in Nigeria for the period beginning in November 2013 and ending in October 2014.Following an investigation into the characteristics of the inflation series (which included an examination of the model's residuals), it was determined that the SARIMA (1, 1, 1) (0, 0, 1)12 model provided the most accurate representation of the inflation rate. A downward trend in inflation was indicated by the projections made over a period of 12 months, beginning in November 2013 and ending in October 2014. The recommendation that was made based on the outcomes of the study's forecast was to provide assistance to the decision-makers in Nigeria.

(Erdoğdu & Yorulmaz, 2019) examined the effectiveness of three models for predicting tax revenue in Turkey from 2006 to 2018. In this investigation, three alternative time series forecasting methods; BATS (Seasonal Box-Cox Transformation, ARMA Errors, Trend, and Seasonal Components), SARIMA (Seasonal Autoregressive Integrated Moving Average), Random Walk; were employed. They split the data set into training and testing at the outset of the analysis. From 2006 to 2014, there was a training phase, and from 2015 to 2018, there was a testing phase. Each forecasting model predicted 36 months based on the outcomes of

ME, RMSE, MAE, MPE, MAPE, MASE, and Theil's U. They discovered that employing the BATS model, as opposed to the conventional SARIMA model, produced better accurate estimates of Turkey's monthly tax revenue.

(Micheni Nelson Kirimi et al., 2022) sought out to find an efficient Holt-Winters and SARIMA models that could be used to forecast Domestic tax revenues in Kenya. Their study utilized the Domestic tax revenues collected in Kenya between Jan 2015 to December 2020.) SARIMA and Holt-Winters time series forecasting methods were applied to the revenue data collected.

SARIMA (0,1,1) $(0,1,1)_{12}$ model was found to be the best model since it had the least Bayesian Information Criterion (BIC=1236.49) and the least forecasting errors (MAPE=6.9, MASE=0.37). The multiplicative Holt-Winters method was slightly superior to the additive method due to its lower error (MAPE=7.43). The study later recommended the use of the two models to forecast Domestic taxes in Kenya and can be used to capture the Domestic taxes revenues with high precision.

(Streimikiene et al., 2018) studied the effects of indirect taxes on the working class and used three distinct time series approaches to anticipate Pakistan's tax collection for the fiscal year 2016 - 2017. The study also examined the effectiveness of three distinct time series models, including the vector autoregression (VAR) model, the autoregressive integrated moving average (ARIMA), and the autoregressive model (AR with seasonal dummies). They concentrated on forecasting for 2017 and used a dataset that ran from July 1985 to December 2016 (monthly). They used tax revenue components such direct tax, sales tax, federal excise duty, and customs duties to anticipate overall tax revenue. According to the study's findings, the ARIMA model provides a more accurate projection of Pakistan's overall tax revenue.

2.3. Review studies on LSTM models

Recurrent neural networks (RNNs) of the Long Short-Term Memory (LSTM) variety are very effective at forecasting time series, including tax income. Forecasting future tax collections requires the accurate modeling of long-term dependencies in time series data, which is where LSTMs come in.

(Rhanoui et al., 2019) contrasted the ARIMA random walk model and the LSTM neural network model on financial time series data for accuracy and efficiency. Additionally, the traditional model and the machine learning model were contrasted. The purpose of this study, which used data for a country's annual budget from 1976 to 2016, was to determine which model could predict the annual budget the most accurately. After making the data stationary, they applied the ARIMA model using several settings, keeping the ARIMA model as the best one (0,1,0). ARIMA and LSTM had RMSE values of 0.239 and 0.222, respectively. The MSE results were 0.057 for ARIMA and 0.049 for LSTM. ARIMA was 0.139 and LSTM was 0.119 in the MAE results. This study demonstrated that while the ARIMA model yields acceptable results, the LSTM model outperforms the ARIMA model's performance. One of the strongest models for time series prediction is the LSTM recurrent neural network because it can recognize non-linear features in financial time series.

(R. Zhang et al., 2022) Using data from 2013 to 2018 with daily, weekly, and monthly datasets, two types of forecasting models (LSTM and ARIMA) with rolling forecast and without rolling forecast at various time scales were created and compared for the incidence of hemorrhagic fever in China for the year 2019. The forecasting performance in 2019 shown that, in rolling forecasting models, LSTM outperformed ARIMA for daily forecasting while ARIMA outperformed LSTM for weekly and monthly forecasting. In 2019, the models that included rolling predictions had lower mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE) values than the direct forecasting models for both ARIMA and LSTM. Rolling forecast results; Daily, RMSE result were, ARIMA – 13.01 and LSTM – 8.05. The MAPE result were, ARIMA – 58.63 and LSTM – 35.70. The MAE result was, ARIMA – 10.06 and LSTM – 5.75.
Weekly, RMSE result were, ARIMA – 31.09 and LSTM – 35.98. The MAPE result were, ARIMA – 11.85 and LSTM – 17.81. The MAE result was, ARIMA – 20.56 and LSTM – 26.21. Monthly, RMSE result were, ARIMA – 108.38 and LSTM – 247.53. The MAPE result were, ARIMA – 8.51 and LSTM – 34.2. The MAE result was, ARIMA – 72.67 and LSTM – 224.42.

2.4. Review studies on MLR models

Machine learning techniques for projecting tax income have drawn more attention in recent years. These techniques have shown to be successful in capturing the intricate connections between economic factors and tax collection. One particular technique used in machine learning to determine the relationship between a number of different variables and a target variable is multiregression.(X. Zhang et al., 2019).

(Petrovski et al., 2015) employed multiple linear regression model to estimate bid price for a company. "Choosing a bid price for contracting and becoming a project stakeholder is a company's key management choice because bid process involves many parties, each with their own interests. As a result, it's crucial to have a bidding model that can forecast the price of the bid in light of evolving threats to the bidding process. Given that MLR uses actual data from the field and had a MAPE (Mean Absolute Percentage Error) of roughly 3% and an R2 = 0.88167 coefficient of determination, it was remarkably accurate. This represents a major advancement above conventional models, which normally have MAPE around 25% and, in some recent studies, MAPE around 19%. The decision-making process can be aided by the proposed model during the competitive bidding process.

(Alashari et al., 2022) evaluated the accuracy and effectiveness of multivariate linear regression (MLR) and vector autoregression (VAR) time series models in estimating the annual maintenance costs of EPDM roofing systems. For a 23-year span, from 1997 to 2019, the study used historical data from 16 distinct EPDM roofing systems. The results of this performance research show that the stepwise MLR model had a little higher average accuracy (85%) than the VAR model in estimating the annual maintenance expenses of EPDM roofs (83%).

2.5. Forecasting and error accuracy

Making predictions about future circumstances or events based on historical and current data is the process of forecasting. In order to forecast trends and make decisions, it is frequently used in business, economics, and finance.

How accurately the anticipated values match the actual values is referred to as error accuracy in forecasting. It is typically expressed as a decimal between 0 and 1 or as a percentage. Mean absolute percentage error (MAPE), mean absolute error (MAE), mean squared error (MSE) are often used indicators of predicting error accuracy. The forecast is thought to be more accurate the lower the error.

Where n is the size of the dataset, A_i is the actual value, F_i is the forecasted value, $\overline{A}_i$ is the mean absolute value, and p is the number of features/predictors.

2.5.1. What is Time Series

A time series is a collection of data points taken at regular periods of time. Time series data is often used to evaluate trends and patterns across time in industries such as finance, economics, and engineering. Time series data can be continuous (for example, temperature) or discrete (for example, stock prices) (e.g. number of sales, number of website visitors).

Time series analysis is the process of modeling and analyzing time series data using statistical techniques in order to extract relevant insights and generate projections. Time series analysis seeks to comprehend the underlying structure and patterns in data, identify trends and seasonality, and make accurate forecasts about future values.

There are several techniques used in time series analysis, including:

- Decomposition: is the process of dividing a time series into its constituent pieces, such as trend, seasonality, and residuals.
- Smoothing: is the process of reducing noise from data in order to expose underlying patterns.
- Forecasting: is the process of predicting future values based on past data.
- Analysis of Trends and Seasonality: Identifying and quantifying the underlying trend and seasonality patterns in data.
- ARIMA: which is a prominent method for modeling and forecasting time series data.

Organizations can use time series analysis to make informed decisions and prepare for the future with more accuracy (Wilson, 2016).

2.5.2. Mean squared error (MSE)

A frequently used metric for comparing anticipated and actual values is mean squared error, or MSE. The average of the squared differences between the predicted and actual values is used to calculate it. It is a common error metric for regression problems, MSE is defined in Equation (1).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(A_i - F_i)^2 \qquad (1)$$

The benefit of MSE is that it penalizes large errors more severely than tiny errors, which is advantageous when attempting to lessen the influence of outliers. It is crucial to scale the data before employing MSE due to its sensitivity to the data's size. Additionally, because it is sensitive to the presence of outliers, the error may appear larger than it actually is (James et al., 2021).

2.5.3. Mean absolute error (MAE)

The mean absolute error (MAE) is a measurement of the difference between predicted and actual values. It is determined by averaging the absolute differences between the forecasted and actual values. MAE is defined in Equation (2).

$$MAE = \frac{1}{n}\{\sum_{i=1}^{N}|A_i - F_i|\} \qquad (2)$$

MAE is a regularly used measure of forecast quality; it is a measure of the average magnitude of errors in a series of forecasts, without taking into account their direction. It demonstrates how inaccurate the projections are.

The advantage of MAE over other measures such as Mean Squared Error is that it is less sensitive to outliers (MSE). Large errors are not substantially penalized because they are not squared. It also uses the same unit of measurement as the data, making it simple to interpret.

However, unlike the Mean Absolute Percentage Error, it does not reveal if the forecast is exceeding or underestimating the actual values (Shumway & Stoffer, 201).

2.5.4. Mean absolute percentage error (MAPE)

The percentage difference between the predicted and actual values is known as the Mean Absolute Percentage Error (MAPE). It is determined by arithmetically averaging the absolute percentage differences between predicted and actual values. MAPE is defined in Equation (3).

$$MAPE = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{|A_i - F_i|}{A_i}\right) \times 100 \qquad (3)$$

MAPE is a commonly used measure of the quality of a forecast, it is a measure of the average magnitude of the errors in a set of forecasts, expressed as a percentage. It gives an idea of how wrong the forecasts are in percentage terms.

MAPE has the advantage of being in the same unit as the data, which makes it easy to interpret. It also allows for comparison between forecasts of different variables or even different forecast models.

However, it has the disadvantage of being undefined when actual values are zero, it also can be sensitive to the presence of outliers, as well as to the scale of the data.

Remember that if your actual values are 0, the formula is not defined, and you'll need to make some changes or switch to a different error measure (Hyndman & Koehler, 2006).

2.5.5. Coefficient of determination (R^2 score)

R-squared (R^2) is a statistical term that shows the proportion of the variance in the dependent variable that can be predicted by the independent variable (s). R-squared is defined in Equation (4).

$$R^2 \ = 1 - \left\{ \frac{\sum_i^n (A_i - F_i)^2}{\sum_i^n (A_i - \bar{A}_i)^2} \right\} \tag{4}$$

R-squared is a popular metric for assessing how well a model fits a set of data. The model fits the data more accurately when the R-squared value is larger. However, because it ignores the model's complexity or if the model is overfitting the data, it cannot tell whether the model is excellent or bad.

It is crucial to keep in mind that R-squared only evaluates how well a model fits the data, not how well it can predict the future (Hastie et al., 2009; Hyndman & Koehler, 2006).

2.5.6. Adjusted Coefficient of determination (Adj-R^2 score)

R-squared has been changed to include an adjustment for the number of independent variables in a model, known as adjusted R-squared (adj-R2). It is a statistical measure that, after accounting for the number of variables, shows what fraction of the variance in the dependent variable can be predicted from the independent variable(s). It has a range of 0 to 1. Adjusted R-squared is defined in Equation (5).

$$R^2_{a\,\mathrm{d}j} = \frac{1-(1-R^2)(n-1)}{n-p-1} \tag{5}$$

When comparing models with various numbers of independent variables, adjusted R-squared is used to adjust the R-squared value for the number of predictors. When comparing models with various numbers of predictors, it is a better indicator of the goodness of fit of a model because it penalizes models with more variables.

It is crucial to keep in mind that, similar to R-squared, adjusted R-squared simply assesses how well the model fits the data and not the model's capacity for making precise predictions (Valerie Watts, 2022).

2.5.7. Root mean squared error (RMSE)

RMSE, or root mean squared error, is a regularly used indicator of the difference between predicted and actual values. It is calculated by taking the mean squared error's square root (MSE). RMSE is defined in Equation (6).

$$RMSE = \sqrt{\frac{1}{n}}\{\sum_{i=1}^{n}(A_i - F_i)^2\} \qquad (6)$$

The average magnitude of the error is measured by the RMSE, which is a commonly used metric of forecast quality. The fact that RMSE is in the same unit as the data makes it simple to interpret. It also penalizes huge errors more than tiny errors, which is useful when reducing the influence of outliers. However, it, like MSE, can be sensitive to the presence of outliers and the scale of the data (Hyndman & Koehler, 2006; James et al., 2021).

2.5.8. Granger Causality

A statistical concept called Granger causality aids in determining the causal relationship between the two variables in a time series of data. It is founded on the hypothesis that if a variable X influences a variable Y, then previous values of X should have knowledge that improves predictions of future values of Y when compared to using only past values of Y.

To put it another way, Granger causality aids in determining if one time series can forecast another better than simply relying solely on its past values. It's crucial to remember that Granger causality does not indicate traditional causality but rather a statistical link between two variables.

Granger causality, as a result, does not prove causality, but rather suggests the presence of a causal relationship between two variables based on statistical evidence. It is also worth mentioning that other factors may influence the link between the variables, which Granger causality analysis may not capture.

2.6. Conclusion

Every country's economy benefits from making the most accurate revenue projections possible since doing so results in a more equitable distribution of future budgets. Time series models have proven to be effective ways to forecast tax revenues from the aforementioned literature analysis, and they can be used for both larger and smaller tax kinds. SARIMA models are a particular method for identifying patterns in data that are both seasonal and non-seasonal. According to studies that have been analyzed, the SARIMA model performs well in estimating sales tax collections as well as personal and corporate income tax receipts. It should be emphasized that although though SARIMA models have been shown to be useful for predicting tax receipts, they might not be appropriate for all taxes or in all circumstances. To produce more precise projections, it's critical to assess the applicability of a given model and combine it with other models and methodologies. In addition, other models, like econometric models and judging approaches, may offer extra perspective and boost predicting accuracy. LSTM is a promising method for estimating tax revenues, according to a study of the literature on the subject. These studies have repeatedly demonstrated that LSTM outperforms conventional statistical models like ARIMA in terms of forecasting accuracy. Additionally, it has been demonstrated that LSTM outperforms other machine learning models like Random Forest and XGBoost. Numerous of this research used various datasets and locations, but they all came to the same conclusion: LSTM is an effective tool for predicting tax income. According to the studies examined, multivariate regression models seem to be useful for predicting tax income. These models' independent variables may incorporate economic statistics like the GDP, inflation, and unemployment rate. These studies have shown that multivariate models perform better in terms of predicting accuracy than univariate models, which may be useful in anticipating tax income. However, a recorded quantitative data sample of the same historical course is required. Because the accuracy of these methods decreases over time, they are more accurate for short-term projections (two to three years). To obtain or generate more accurate long-term forecasts, more information about the variables of interest must be gathered, the model must be well-defined, and statistics such as root mean squared error, mean absolute percentage error, Akaike information criterion, quadratic lost function, and many others must be taken into consideration. In addition, it is necessary to monitor the forecasts that were obtained to determine whether any adjustments are necessary. The theory of SARIMA, LSTM, and MLR models is illustrated in the following chapter by examining model specification and test statistics for forecasting.

CHAPTER 3 – PROPOSED METHODOLOGY

3.1. Introduction

This chapter examines the theory underlying the Seasonal autoregressive integrated moving average (SARIMA), Long short-term memory networks (LSTM), and multivariate linear regression (MLR) models. In section 3.5, we conclude our discussion of forecasting by describing the criteria we will use to evaluate the accuracy of the model's prediction.

3.2. SARIMA modeling

SARIMA (Seasonal Autoregressive Integrated Moving Average) is a type of time series forecasting model that accounts for both seasonality and autocorrelation. The SARIMA model is typically denoted as:
ARIMA (p, d, q) (P, D, Q) s

The Formula for SARIMA is defined in Equation (7);

$$y\_t = c + \phi\_1y\_t - 1 + \ldots + \phi\_py\_t - p - \theta\_1e\_t - 1 - \ldots - \theta\_qe\_t - q + (1 - B)\hat{}d * (1 - B\_s)\hat{}D * (y\_t - c\_s) + e\_t + \Phi\_1 * (1 - B)\hat{}d * (1 - B\_s)\hat{}D * (y\_t - s - c\_s) + \ldots + \Phi\_P * (1 - B)\hat{}d * (1 - B\_s)\hat{}D * (y\_t - ps - c\_s) - \Theta\_1e\_t - s - \ldots - \Theta\_Qe\_t - Q \tag{7}$$

- y_t is the dependent variable at time t
- c is a constant term
- ϕ_1, ..., ϕ_p are the autoregressive coefficients for the non-seasonal component
- y_t-1, ..., y_t-p are the lagged values of the dependent variable for the non-seasonal component
- θ_1, ..., θ_q are the moving average coefficients for the non-seasonal component
- e_t-1, ..., e_t-q are the lagged values of the error term for the non-seasonal component
- (1-B)^d and (1-B_s)^D are the non-seasonal and seasonal difference operators of orders d and D, respectively
- c_s is the seasonal constant term
- Φ_1, ..., Φ_P are the autoregressive coefficients for the seasonal component
- y_t-s, ..., y_t-ps are the lagged values of the dependent variable for the seasonal component

- Θ_1, ..., Θ_Q are the moving average coefficients for the seasonal component
- e_t-s, ..., e_t-Q are the lagged values of the error term for the seasonal component
- s is the seasonal period.

3.2.1 Autoregressive (AR) component of (p)

The dependence between an observation and several lag observations is modeled by the autoregressive (AR) component of a SARIMA model. A time series' current value is a linear combination of its previous values, with the coefficients of the linear combination being specified by a set of parameters. This is the core notion behind an AR model. The first-order autoregressive model, or AR (1) model, is a straightforward illustration of an AR model and is defined in Equation (8).

$$Y\_t = c + \phi * Y\_t - 1 + \varepsilon\_t \qquad (8)$$

Where Y_t is the current observation, Y_t-1 is the previous observation, c is a constant, ϕ is the autoregressive coefficient, and ε_t is the error term. The autoregressive coefficient, ϕ, is a scalar value between -1 and 1, is a measure of how closely the current observations relate to earlier ones. Strong positive relationships are indicated by a value ϕ that is close to 1, whereas strong negative relationships are indicated by a number **φ** that is close to -1.
Maximum likelihood estimation (MLE) or least squares estimation (LSE) are frequently used to estimate the autoregressive component of a SARIMA model, and the Yule-Walker equations or the Burg method are frequently used to estimate the autoregressive coefficients. Although the SARIMA model's autoregressive (AR) component is a potent tool for modeling time series data, it can be difficult to precisely estimate the model's parameters. To make sure that the model is a good fit for the data, it is crucial to examine the model residuals for normality and independence (WALTER ENDERS, 2015).

3.2.2. Degree of differencing (d)

The differencing component (d) of a SARIMA model is used to make a non-stationary time series stationary by removing the trend component from the data. It is defined as the number of times that the original time series is differenced to obtain a stationary series. For example, if d=1, then the first differences of the original time series are taken. If d=2, then the first differences of the first differences are taken. The differencing component (d) can be represented mathematically as the operator Δ^d, where Δ is the difference operator and d is the degree of differencing. For example, a time series Y_t can be transformed into a stationary time series Δ^dY_t by taking the d-th order differences of Y_t, the equation is defined in Equation (9).

$$\Delta^{d} Y\_t \;=\; Y\_t \;-\; Y\_t - d \qquad (9)$$

3.2.3. Order of the moving average (MA) component (q)

Moving average (MA) is used to model the dependence between an observation and the residual error from a moving average model applied to lagged observations. The fundamental concept underlying an MA model is that the current value of a time series is a linear combination of the error term from a moving average model applied to lagged observations, with the coefficients of the linear combination being determined by a set of parameters.

A straightforward example of an MA model is the MA (1) model, which defined in Equation (10).

$$Y\_t \;=\; \mu \;+\; \varepsilon\_t \;+\; \theta\varepsilon\_\{t-1\} \qquad (10)$$

Where Y_t is the current observation, μ is the mean of the series, ε_t is the error term, and θ is the moving average coefficient. The moving average coefficient, θ, is a scalar value between -1 and 1, and represents the strength of the relationship between the current observation and the past error term. Strong positive relationships are indicated by a value ϕ that is close to 1, whereas strong negative relationships are indicated by a number ϕ that is close to -1.

The moving average component of a SARIMA model is typically estimated using least squares estimation (LSE) or maximum likelihood estimation (MLE).

3.2.4. Seasonal autoregressive (SAR) component (P)

Modeling the relationship between an observation and several lagging seasonal observations can be done with the SAR part of a SARIMA model. The present value of a time series is thought to be a linear combination of its previous seasonal values, with the coefficients of the linear combination being determined by a given set of parameters in a SAR model. The first-order seasonal autoregressive model (SAR (1)) is a simple example of a SAR model. The formula for SAR is defined in Equation (11).

$$Y\_t \;=\; c \;+\; \varphi_{S} Y\_\{t-m\} \;+\; \varepsilon\_t \qquad (11)$$

Where Y_t is the current observation, Y_t-m is the previous observation at the same season, c is a constant, φ_s is the seasonal autoregressive coefficient, m is the number of seasons, and ε_t is the error term. The seasonal autoregressive coefficient, ϕ_s, is a scalar value between -1 and 1, and represents the strength of the relationship between the current and previous observations at the same season Strong positive relationships are indicated by a value ϕ that is close to 1, whereas strong negative relationships are indicated by a number ϕ that is close to -1.

Note that in order to accurately model the seasonal factor, it’s important to have a time series that have enough seasonal data, and to choose the right number of seasons (m) correctly.

3.2.5. Degree of seasonal differencing (D)

To eliminate the seasonal component from a time series and make it stationary, a SARIMA model's degree of seasonal differencing (D) is used. Similar to regular differencing (d), seasonal differencing takes the differences between consecutive observations that occurred during the same season as opposed to between consecutive observations.

The seasonal differencing component is represented mathematically as the operator Δ_m^D, where Δ_m is the seasonal difference operator, m is the number of seasons, and D is the degree of seasonal differencing.

For example, a time series Y_t can be transformed into a stationary time series Δ_m^DY_t by taking the D-th order seasonal differences of Y_t, the equation is defined in Equation (12).

Δ_m^D Y_t = Y_t − Y_{t − m}^D (12)

In order to stabilize the time series' variance and guarantee that the model is a good fit for the data, the degree of seasonal differencing (D) is a crucial parameter in the SARIMA model. It is typical to experiment with different values of D until a stationary series is obtained if a time series is found to have a seasonal component. It's crucial to remember that choosing D has a direct impact on choosing the number of seasons (m), as well as P and Q. When selecting the values of D, P, and Q, careful consideration of the time series and its seasonal pattern is required.

3.2.6. Order of the seasonal moving average (SMA) component (Q)

The dependence between an observation and the residual error from a seasonal moving average model applied to lagged observations is modeled by the seasonal moving average (SMA) component of a SARIMA model. A seasonal moving average (SMA) model works on the fundamental premise that the current value of a time series is a linear combination of the error term from a SMA model applied to lagged observations, with the coefficients of the linear combination being influenced by a set of parameters. The first-order seasonal moving average model, or SMA (1) model, is a straightforward illustration of a SMA model and is defined in Equation (13).

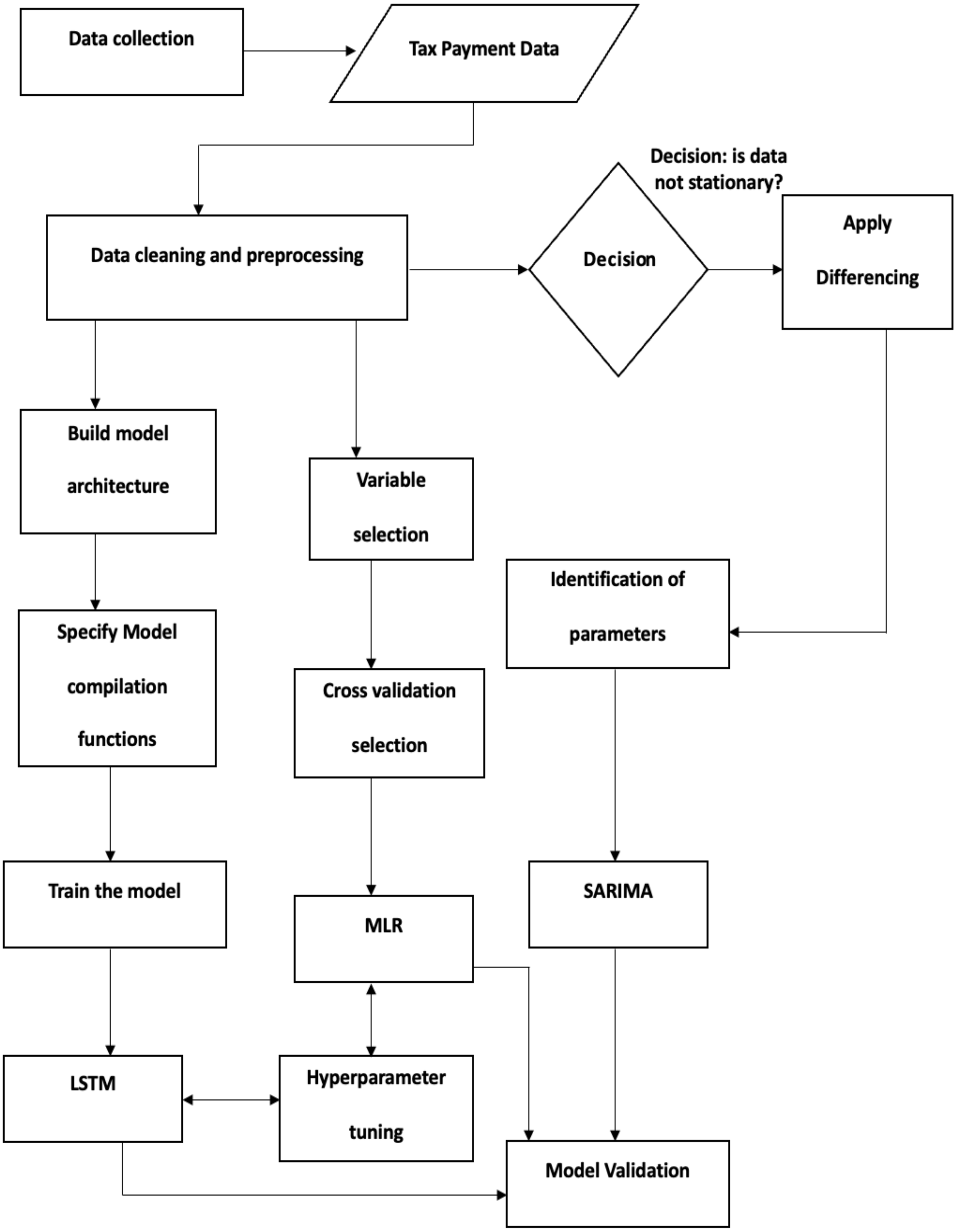
$$Y\_t \;=\; \mu \;+\; \varepsilon\_t \;+\; \theta\_s\, \varepsilon\_\{t-m\} \tag{13}$$

Where Y_t is the current observation, μ is the mean of the series, ε_t is the error term, m is the number of seasons, and θ_s is the seasonal moving average coefficient. The seasonal moving average coefficient, θ_s, is a scalar value between -1 and 1, and represents the strength of the relationship between the current observation and the past error term of the same season. A value of θ_s close to 1 indicates a strong positive relationship, while a value of θ_s close to -1 indicates a strong negative relationship. It is important to note that the selection of Q is closely related to the selection of the number of seasons (m) as well as the selection of P and D.

3.2.7. The process of building a SARIMA model

1. Using a time series data visualization to spot any patterns or trends.
2. Testing the time series data' stationarity.
3. Using methods like the partial autocorrelation function (PACF) and autocorrelation function (ACF), we identify the optimal values for the parameters p, d, and q, as well as the seasonal parameters P, D, and Q.
4. Using metrics like root mean square error (RMSE), mean squared error (MSE) and mean absolute error (MAE) to evaluate the SARIMA model's performance after fitting it to the time series data.
5. Lastly, predicting future values of the time series using the fitted model (WALTER ENDERS, 2015).

Figure 3.1. Flowchart for building the SARIMA, LSTM and MLR model



3.3. LSTM modeling

The Recurrent Neural Network (RNN) known as Long Short-Term Memory (LSTM) is capable of identifying long-term dependencies in sequential data. LSTMs were first presented in a paper by (Sepp Hochreiter & J□urgen Schmidhuber, 1997) in the paper "Long Short-Term Memory". When it comes to processing data, LSTMs rely on a set of gates (input, forget, and output) to regulate the information's progression through the network. The gates are operated by learned weights that determine what data is allowed through, what data is discarded, and what data is outputted.

LSTMs are based on the concept of a cell state that is propagated throughout the network in addition to the input and output. One way to think of the cell state as a memory is as a long-term storage mechanism for information. The gates in LSTMs regulate the entry and exit of data into and out of the cell state, giving the network the ability to pick and choose what to keep and what to throw away.

The core components of an LSTM cell are:

1. Input Gate (i): regulates how much data enters the current state of the cell.
2. Forget Gate (f): controls the amount of information that is forgotten from the previous cell state.
3. Output Gate (o): regulates the rate at which data leaves the cell.
4. Cell State (c): holds data that will be transmitted from one time step to the next.
5. Hidden State (h): the value produced by the LSTM cell and fed into the next iteration.

The LSTM model can be defined by a series of Equations (14-18):

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}C_{t-1} + b_i) \quad (14)$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \quad (15)$$

$$c_t = i_t tanh\left(W_{xf}x_t + W_{hc}h_{t-1} + b_c\right) + f_t c_{t-1} \quad (16)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (17)$$

$$h_t = o_t tanhc_t \quad (18)$$

Where x is the input, h is the hidden state, W_{ci}, W_{hc}, W_{co}, W_{cf}, and W_{xi} are the parameters (weights and biases) of the LSTM model, sigmoid and tanh are activation functions and the b terms denote bias vectors.

Additionally, the LSTM model can also have other attributes such as:

1. Number of layers: LSTMs can be stacked on top of each other to create deeper architectures that can learn more complex representations.
2. Number of units: how many neurons are contained within each LSTM cell, which ultimately determines how much data can be stored and processed by the model.
3. Dropout: a regularization technique that randomly drops out units during training to prevent overfitting.
4. Bidirectional: a configuration where the LSTM reads the input sequence in both forward and backward directions.
5. Attention: a method that enables the LSTM to generate output by concentrating on particular segments of the input sequence.

3.3.1. Recurrent Neural Network (RNN)

When dealing with sequential data, like time series or natural language, a Recurrent Neural Network (RNN) is the best artificial neural network to use. RNNs are equipped with a "memory" that allows them to recall and apply information from previously processed inputs to the current input. An RNN's "recurrent unit" is its smallest component; it receives an input and a "hidden state" (the network's memory of previous inputs) and outputs a value and a new hidden state. When one recurrent unit finishes processing data, it passes on the hidden state to the next one in the chain. There are several variations of RNNs, including the **Long Short-Term Memory (LSTM)** and Gated Recurrent Unit (GRU) networks. These variations have been designed to address some of the issues with training traditional RNNs, such as the vanishing gradient problem.

Overall, RNNs are a powerful tool for processing sequential data, and have been used in many state-of-the-art models for natural language processing and other tasks.

3.3.2. The process of building a LSTM model

The input sequences are processed in a step-by-step manner, one time step at a time, by an LSTM model, which then generates predictions. The model generates a new hidden state as well as an output after each time step, using the current input as one of its inputs and the previous hidden state as the other input. After that, the output is sent through an output layer, which maps it to the output space that was specified earlier. Inside of the LSTM cell, the process of generating the output as well as the new hidden state is carried out utilizing the input gates, output gates, forget gates, and cell state(Sepp Hochreiter & J□urgen Schmidhuber, 1997).

The model learns the best settings for the gates' and output layer's parameters (weights and biases) during training by reducing the error between the expected and actual outputs. Once the model has been trained, it may be used to make predictions on fresh input sequences by applying the same time-step-by-time-step processing and learnt parameters that it used during training.

For the purpose of language modeling, for instance, the LSTM receives as input a string of words, one at a time, and at each time step it generates a probability distribution over the vocabulary for the following word. The procedure continues until the end of the input sequence is reached, at which point the algorithm will stop and the word with the highest probability will be chosen as the anticipated next word.

Forecasting time series data using an LSTM model involves several steps:

1. Data preprocessing: Prepare the time series data for use as input by the model by eliminating noise, standardizing the values, and applying any necessary transformations. Typically, one part of the data is used to train the model, while the other is utilized to test and evaluate how well the model performed.
2. Model architecture: The number of input and output variables, time steps, hidden layers, and units are some of the special properties of the time series data that are taken into account when designing the LSTM model architecture.
3. Training: Using a suitable optimization approach, such as stochastic gradient descent, the model is trained on the training set (SGD). In order to reduce the discrepancy between the projected output and the actual output, the model's parameters (weights and biases) are adjusted during the training phase.
4. Forecasting: The trained model can then be used to make predictions on unlabeled data. As part of time series forecasting, the model receives as input a series of past values and outputs a prediction for the next value in the series. The process is iterated as many times as necessary to produce a forecast for the time horizon of interest.
5. Evaluation: The test set is used to assess the model's accuracy by comparing the predicted and actual values and then computing an error measure such as the mean squared error (MSE), mean absolute error (MAE) or root mean squared error (RMSE).

It's worth noting that LSTMs are particularly useful for time series data that has a temporal dependency, where the current value is dependent on previous values. LSTMs can handle this type of sequential data by using its memory cells, gates, and hidden state to process it (Siami-Namini et al., 2019).

3.3.3. Hyperparameter Tuning

Hyperparameter tuning involves finding the optimal values for hyperparameters that are defined by the user before training a machine learning algorithm. These hyperparameters cannot be learned from the data and need to be specified prior to training. In the case of LSTM models, hyperparameters may include the number of layers, number of neurons in each layer, learning rate, batch size, dropout rate, and number of epochs. These hyperparameters can impact the performance of the model, so selecting the optimal values is crucial for achieving the best possible performance. Hyperparameter tuning can be done manually or automatically using techniques such as grid search, random search, or Bayesian optimization. The goal is to optimize the model's performance on a given dataset, leading to better accuracy and predictions.

In summary, forecasting time series data using an LSTM model involves several steps: data preprocessing, model architecture, training, forecasting and evaluation. LSTMs are particularly useful for time series data that has a temporal dependency, where the current value is dependent on previous values.

3.4. MLR modeling

MLR is a type of linear regression that permits several independent variables (also known as predictors or features) to be used to predict a single dependent variable (also known as the response or outcome). The purpose of MLR is to determine the best-fitting line or hyperplane that explains the relationship between the variables and to use this model to forecast the dependent variable's future values. The MLR model is defined in Equation (19).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{19}$$

The dependent variable is Y, the independent variables are X1, X2, Xn, the y-intercept is b0, and the independent variables' coefficients are b1, b2., bn. When all of the other variables are maintained constant, the change in the dependent variable that is represented by these coefficients occurs when there is a change of one unit in the independent variable that corresponds to it.

Finding the values of the coefficients in an MLR model that minimize the gap between the predicted values of the dependent variable and the actual values is one of the steps involved in the process of training an MLR model. This is often accomplished through the use of a technique known as least squares, which seeks to minimize the sum of the squared disparities that exist between the values that were predicted and those that were actually observed.

After the model has been trained, it can be used to produce forecasts based on data that it has not previously encountered. In addition, the coefficients of the variables that are independent can be used to infer the relative importance of each variable in terms of its ability to predict the variable that is dependent (Hastie et al., 2009).

When employing an MLR model, a few assumptions are frequently made. The fact that there is a linear relationship between the independent and dependent variables is among the most significant. In other words, the change in the dependent variable follows the change in the independent variable in a linear fashion. Additionally, it is presumed that the errors are independent, properly distributed, and homoskedastic in nature; these assumptions are typically verified using diagnostic plots and statistical tests (James et al., 2021).

In conclusion, Multivariate Linear Regression (MLR) is a variation of Simple Linear Regression that enables the use of numerous independent variables to forecast a single dependent variable. It entails determining the line or hyperplane that best fits the data and makes the assumptions that the relationship is linear and the errors are independent, normally distributed, and have a constant variance.

3.4.1. The process of building a MLR model

When employing a multivariate linear regression (MLR) model to forecast time series data, historical data on many variables are used to forecast the future values of the target variable. The following steps are often included in the process:

1. Data collection and preparation: Gathering and getting ready the historical data for analysis is the initial phase. This entails preparing the data for modeling by cleaning, formatting, and possibly changing it.
2. Feature selection: The model's most pertinent independent variables (features) are then chosen to be included. This can entail applying methods like principal component analysis (PCA), correlation analysis, or feature importance measurements.

3. Model fitting: Using the chosen characteristics, the MLR model is subsequently fitted to the historical data. Finding the values of the coefficients that reduce the discrepancy between the target variable's anticipated values and its actual values is often required.
4. Model evaluation: Metrics like mean square error (MSE), mean absolute error (MAE), and coefficient of determination are used to assess the effectiveness of the fitted model (R-squared)
5. Forecasting: Once trained, the model can be used to make predictions based on fresh, unused data. To predict future values of the target variable, the model is applied to time series data.
6. Model updating: To increase the forecast's accuracy over time, the model should be revised to incorporate new data as it becomes available.

It's vital to keep in mind the temporal relationships between observations while working with time series data, as well as the possibility of trends or seasonality in the data, which can be handled by methods like differencing, decomposition, or state-space models (Shumway & Stoffer,2017).

CHAPTER 4 – EXPERIMENTAL RESULTS

4.1. Introduction

When presenting experimental results, it is crucial to include information about the dataset that was used. This information helps to provide context and enables others to understand the limitations and generalizability of the results. A typical introduction to the dataset in experimental results may include details such as:

1. The source of the dataset: where it was obtained and any relevant background information.
2. The size of the dataset: how many samples or instances are included in the dataset.
3. The preprocessing steps: any transformations or manipulations applied to the data before analysis.
4. Characteristics of the data: any relevant information about the data, such as the type of data (e.g., text, image, audio), the format of the data, or any unique properties of the data.

By including this information, researchers can provide transparency and increase the credibility of their results, allowing others to understand the data that was used and potentially replicate the experiment.

4.2. SARIMA Performance and Results

This section shows the result and how it was gotten using the SARIMA model and the error scores from 5 different error accuracy scores. The best fit to use with the SARIMA model was:

(1, 1, 1)x(1, 1, 1, 12), this resulted in a better performance than any other combination. We used just the **total revenue** in Table1 in the appendix which shows the full dataset used.

Figure 4.1: Autocorrelation graph of the dataset

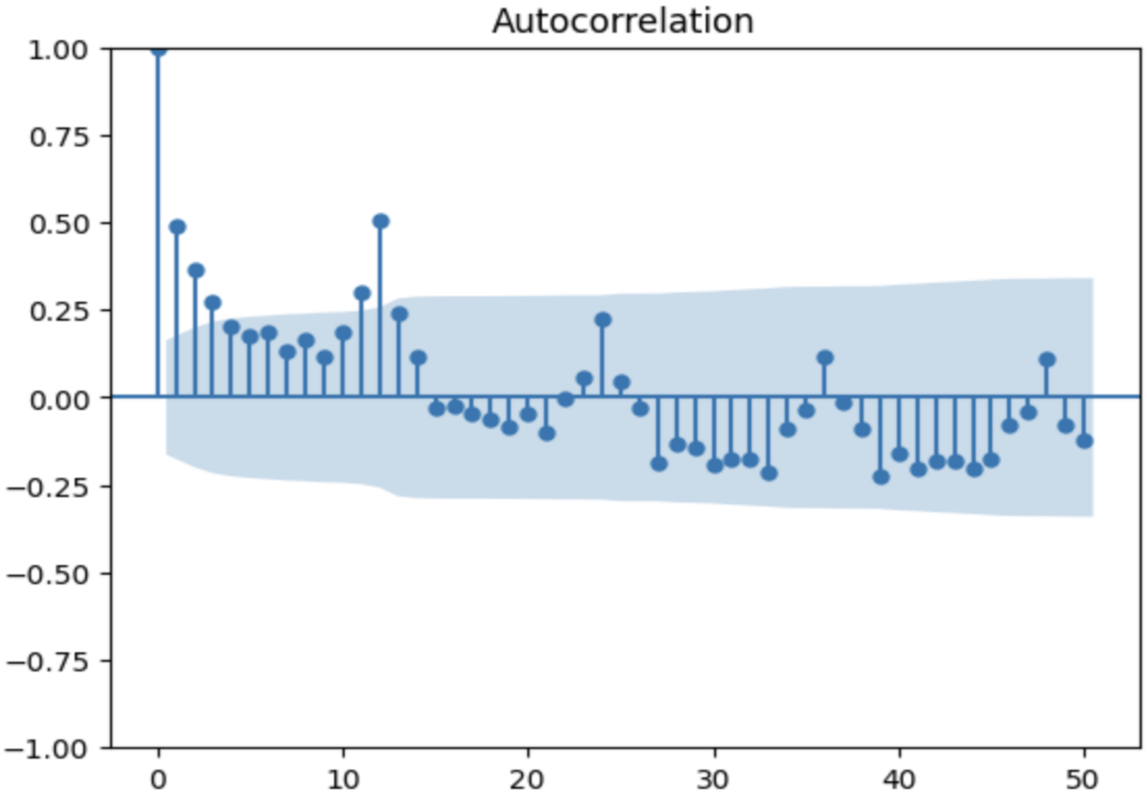


Figure 4.2: Partial autocorrelation graph of the dataset

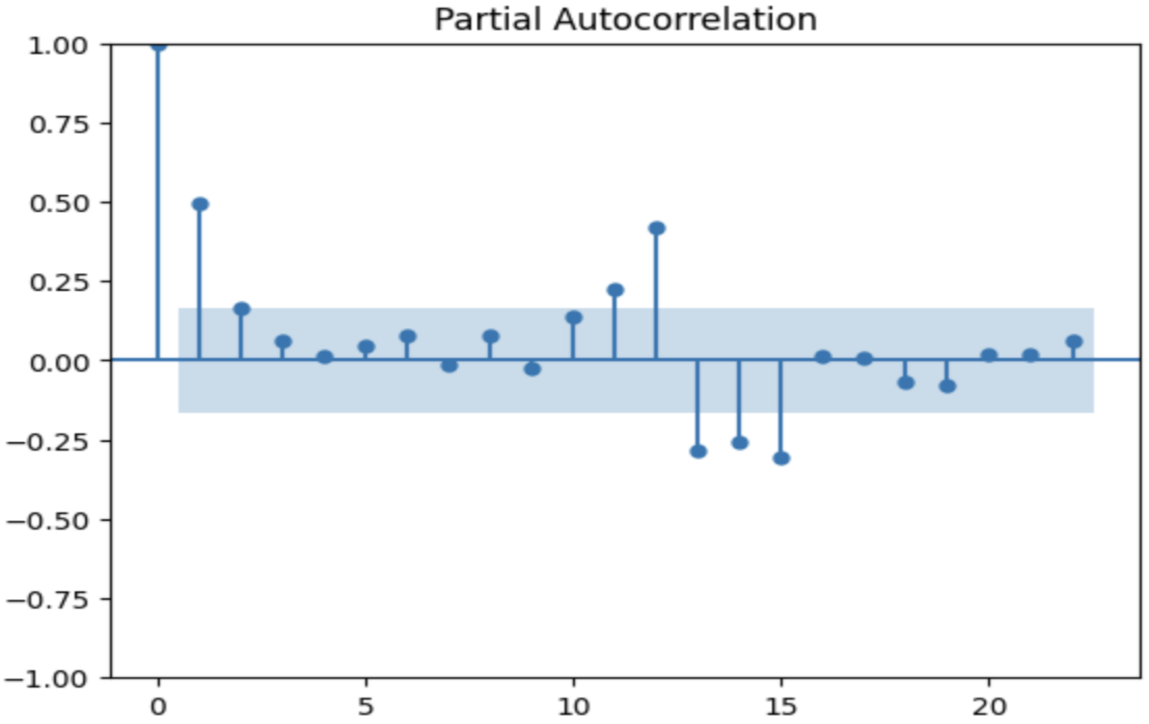
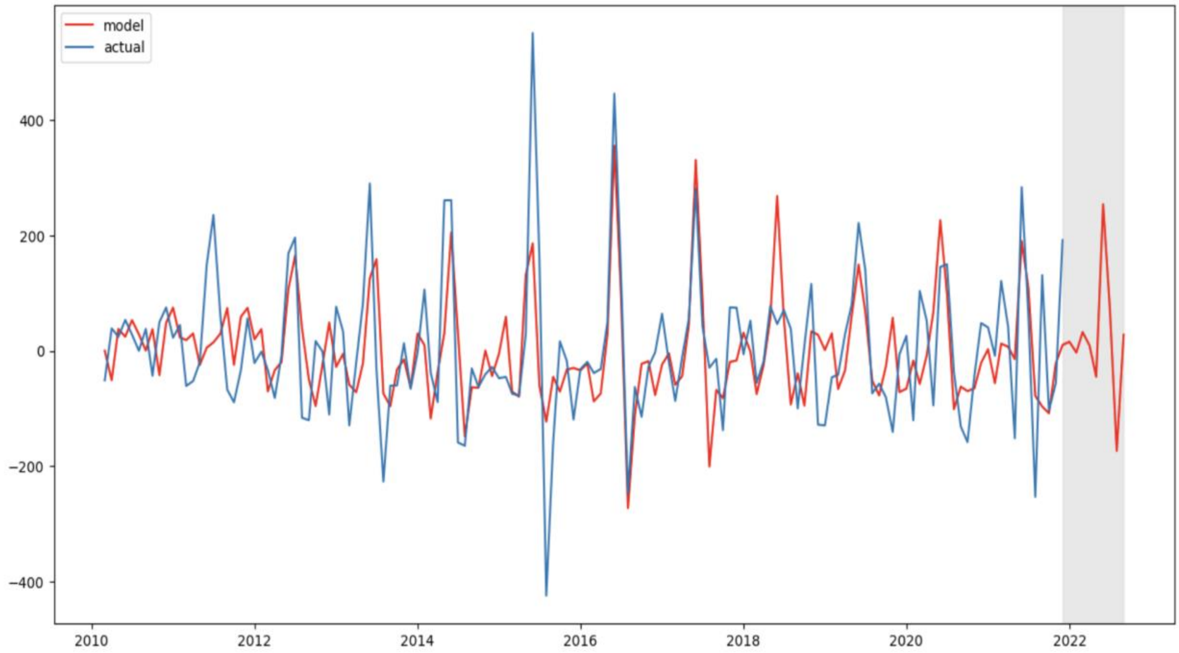


Table 4.1: Error Accuracy Result for SARIMA, LSTM, MLR

Model	K	R2	MSE	MAE	RMSE	MAPE	Adj R2	Diff
SARIMA	-	38.4%	9571.09	74.57	97.83	32.31	-	736.50B
LSTM	-	98.9%	137.48	5.89	11.73	0.016	98.8%	26.4B
MLR	**10**	97.4%	319.07	5.57	17.86	0.014	97.3%	387.5B
MLR	**15**	98.8%	143.67	4.16	11.98	0.0098	98.7%	68.98B
MLR	**20**	97.5%	316.40	5.67	17.79	0.014	97.4%	255.89B

Table 4.1 the SARIMA model has the lowest performance across all evaluation metrics. It has a significantly lower R2 value (38.4%) compared to the other models, as well as higher MSE, MAE, RMSE, and MAPE values. Additionally, the SARIMA model lacks information on the adjusted R-squared value.

Figure 4.3: graph showing the differences between the actual and the predicted value for the SARIMA model



According to Figure 4.3, the SARIMA model performed poorly in limiting the differences between the actual and the predicted value. We came to this conclusion not only because of how the graph above looks like but based on the sum of the differences between the actual and predicted value resulted in **736.50B**

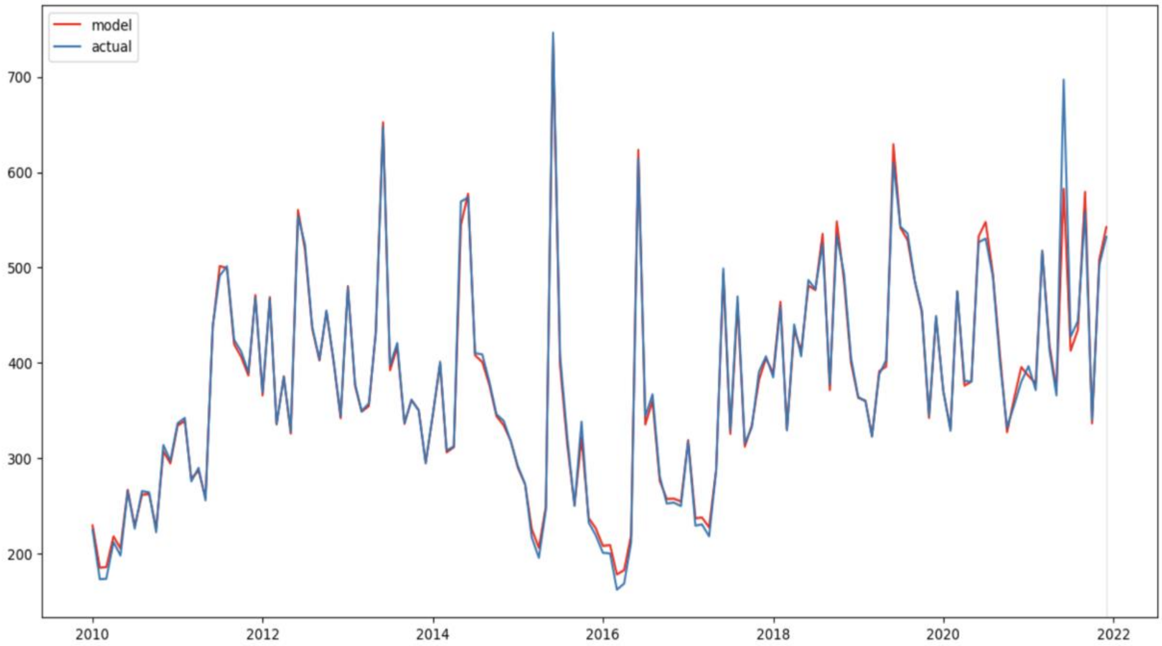
4.3. LSTM Performance and Results

This section shows the result and how it was gotten using the Multivariate LSTM model and the error scores from 6 different error accuracy scores, we included the adjusted r2 score here since we are dealing with multiple independent variables. We used just the **total revenue** as the dependent variable and CIT, VAT, PPT, CGT, EDT, SD, CONS, NITDEF as the independent variables to predict the dependent variable which can be seen in **Table1** in the appendix which shows the full dataset used. I used the mean squared error loss function and the Adam optimizer for training. The model consists of a single LSTM layer with 64 memory units. We used just the **total revenue** in Table1 in the appendix which shows the full dataset used.

Table 4.1 the LSTM model has the highest R2 value (98.9%), indicating a strong correlation between the predicted and actual values. It also has the lowest MSE, MAE, RMSE, and MAPE values, suggesting the best predictive performance among the listed models. The LSTM model also has a high adjusted R-squared value (98.8%), indicating a good balance between model complexity and fit.

According to Figure 4.4, the LSTM model performed way better than the SARIMA model which was all over the place especially the MAPE score which was incredibly high for a model and was slightly better than the MLR model. We came to this conclusion not only because of how the graph above looks like but based on the sum of the differences between the actual and predicted value resulted in **26.4B.**

Figure 4.4: graph showing the differences between the actual and the predicted value for the LSTM model



4.4. MLR Performance and Results

This section shows the result and how it was gotten using the MLR model and the error scores from 6 different error accuracy scores, we included the adjusted r2 score here since we are dealing with multiple independent variables. We used just the **total revenue** as the dependent variable and CIT, VAT, PPT, CGT, EDT, SD, CONS, NITDEF as the independent variables to predict the dependent variable which can be seen in **Table1** in the appendix which shows the full dataset used. We used a cross validation (K) of 15 because it gave the best r2 score and sum of differences between actual and predicted values. The tables below show the scores of 3 different K values that was tested.

The MLR models with K = 10 and K = 15 have similar accuracy to the LSTM model. However, they are less stable, as evidenced by their higher Diff values.The MLR model with K = 20 is less accurate than the other models. This is likely because it is overfitting the data.

Figure 4.5: graph showing the differences between the actual and the predicted value for the MLR model where K = 10 and 20

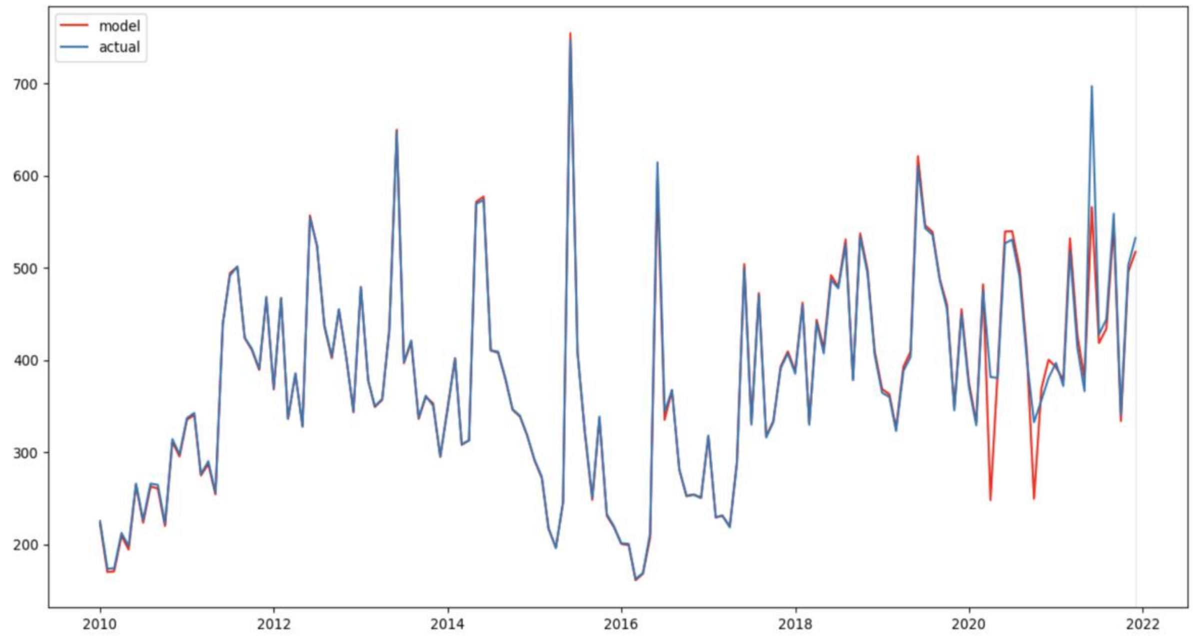
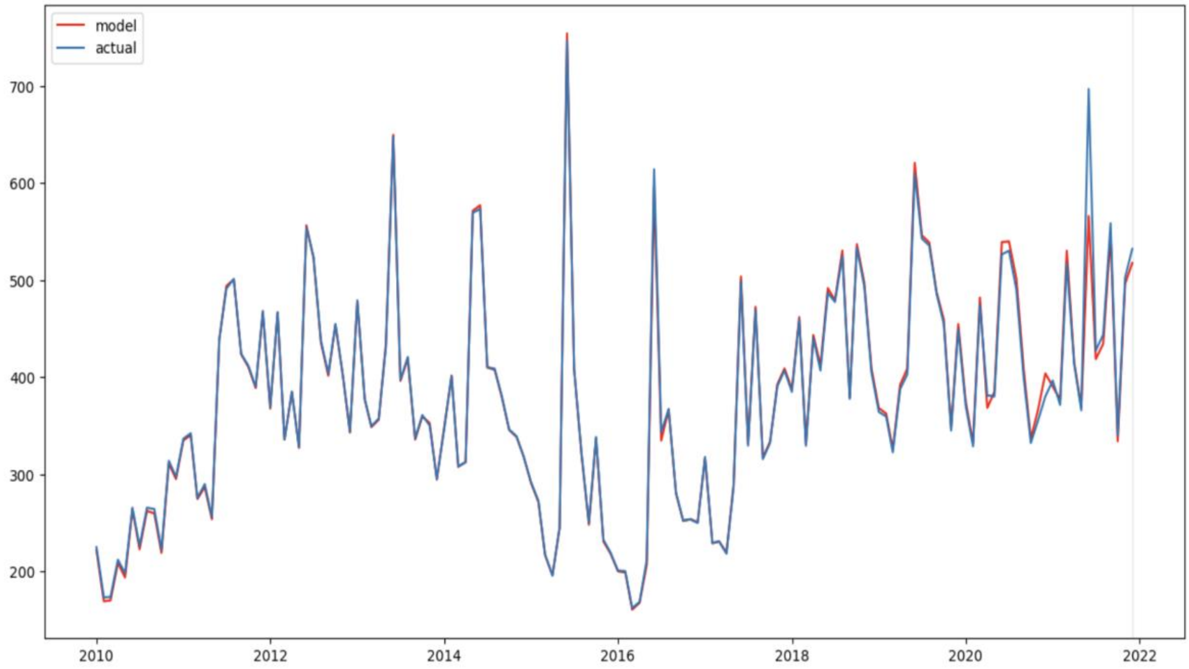


Figure 4.6: graph showing the differences between the actual and the predicted value for the MLR model where k = 15



According to Figure 4.6, the LSTM model and the MLR model with cv of 15 perform very well in predicting the target variable. The LSTM model has a slightly higher R2 value and lower MSE, indicating a better fit to the data and more accurate predictions overall. However, the MLR model with cv of 15 has a slightly lower MAE, suggesting it performs slightly better in terms of average prediction errors. We came to this conclusion not only because of how the graph above looks like but based on the sum of the differences between the actual and predicted value resulted in **68.98B.**

4.5 Granger Casual relationship between the Dependent and Independent

In order to investigate the Granger causal relationship that exists between dependent and independent variables, a Granger Causality Test was carried out. The conclusion drawn from the Granger Causality Test in Table 4.3 suggests that there is evidence of a uni-directional causal relationship between Petroleum Profit Tax and Total Tax Revenue, as well as between Capital Gain Tax and Total Tax Revenue. This means that changes in Petroleum Profit Tax and Capital Gain Tax can help predict changes in Total Tax Revenue, but the reverse is not true. Additionally, there is evidence of a uni-directional causal relationship between Total Tax Revenue and Value Added Tax.

Table 4.2: Granger Causality Test

Null Hypothesis:	**Probability**
TTR does not Granger Cause CIT	**0.4281**
CIT does not Granger Cause TTR	**0.8022**
TTR does Granger Cause VAT	**0.0117(*)**
VAT does not Granger Cause TTR	**0.6039**
TTR does not Granger Cause PPT	**0.7578**
PPT does Granger Cause TTR	**0.0492(*)**
TTR does not Granger Cause CGT	**0.2649**
CGT does Granger Cause TTR	**0.0402(*)**

Decision: The asterisk (*) denote the rejection of the null hypothesis.

CHAPTER 5 – CONCLUSION AND DISCUSSION

5.1. Discussion

According to the findings and based on the Table 4.1, it appears that the LSTM model has the best performance. It achieved a high R2 value of 98.9%, indicating a strong correlation between the predicted and actual values. Additionally, it has the lowest MSE, MAE, RMSE, and MAPE values among multivariate linear regression model (MLR) and seasonal autoregressive integrated moving average (SARIMA). The LSTM model also has a high adjusted R-squared value of 98.8%, which suggests that it has a good balance between model complexity and fit. Therefore, based on the given evaluation metrics in Table 4.1, the LSTM model seems to be the best choice among the options provided. The study's findings also indicated that indirect taxes account for the majority of tax revenue (PPT, CIT, VAT, CONS, EDT, NITDEF). The outcomes are consistent with earlier studies by (Aamir et al., 2011; Chaudhry & Munir, 2010; Himani, 2016) , they examined the factors of tax collection in India and Pakistan and concluded that indirect taxes play a key role in tax collection in Pakistan. They concluded that indirect taxes have increased the gap between the wealthy and the poor and further exploit the working class's susceptibility. In our research, the total tax income was the dependent variable, and direct (CGT and SD) and indirect (PPT, CIT, VAT, CONS, EDT, NITDEF) taxes were the independent variables. Earlier research used the same variables. (Aamir et al., 2011; Himani, 2016; Luqman, 2014; Qadirpatoli et al., 2012). They came to the same conclusion as the previous group, which was that indirect taxes are more prevalent than direct taxes, with the exception of the findings of (Himani, 2016) which he came to the conclusion that in the case of India, total tax revenue is more notable and dominant than direct taxes. The capacity to incorporate additional variables for prediction makes MLR effective, but we neglected to take into account other macro-economic factors that would have strengthened forecasting and brought it much closer to reality. The study also demonstrated that the PPT tax has a Granger effect on overall tax income. These results are very much in line with the true picture or ground realities and consistent with earlier research studies such as (Akhter & Hassan, 2012; Eugene & Abigail, 2016; Karagöz, 2013) because as a nation that produces oil, Nigeria has a tax rate of 50 percent for petroleum operations that are covered by production sharing contracts, 65.75 percent for petroleum operations that are not covered by PSCs, and 85 percent for petroleum operations that are not covered by PSCs after the first five years.

5.2. Conclusion

The purpose of this study is to analyze the performance of three predicting tax revenue models for Nigeria from 2010-01 to 2021-12 and identify the most effective model. In this investigation, three distinct time series forecasting methodologies, including the multivariate long short-term memory networks (LSTM), seasonal autoregressive integrated moving average (SARIMA) and multivariate linear regression (MLR) models, were employed. The results of the study revealed the efficiency of three distinct time series models, as well as the precise outcomes of forecasting total tax income for the preceding years in order to determine if it was adequate, so laying the right groundwork for policymaking by the Nigerian government.

The LSTM model outperforms the MLR slightly though and not on all metrics used (MSE, R2 Score, RMSE) but completely outperforms SARIMA models according to all evaluation metrics (MSE, R2 Score, RMSE, MAE, MAPE). We discovered that forecasting Nigeria's monthly tax revenue series using the LSTM model rather than SARIMA or MLR produced more accurate forecasts because LSTM models are specifically designed to handle sequential or time-series data. They have the ability to capture and learn from temporal dependencies and patterns in the data, which is crucial in many time-series prediction tasks. MLR models, on the other hand, assume independence between data points and do not consider the sequential nature of the data. According to the LSTM model, we were able to determine that there is no substantial discrepancy between the actual total tax revenue and the predicted tax revenue. This study's empirical findings aid specialists in the process of preparing government budgets. The LSTM model may be superior for additional forecasting of other tax types, such as Petroleum Profit Tax (PPT), Value-Added Tax (VAT), and Companies Income Tax (CIT). Furthermore, the empirical findings of this work will be used to develop combination forecasting models (Jabeur et al., 2022).

It would have been more efficient to use more realistic independent variables like Real GDP Rate, crude oil production, Nigeria stock exchange, Employment Population, Unemployment Rate and Inflation and exchange rate to get better results. In the future we should be considering Hybrid models, combinations of two or more models, that will complement each other strength and weakness for a more robust and efficient prediction or forecasting (Bala & Shuaibu, 2022).

5.3. Applications

For governments and tax authorities to efficiently manage and distribute resources, forecasting tax revenue is a critical duty. Forecasting tax income accurately enables decision-makers to make well-informed choices regarding public spending, taxation, and policy.

Here are some possible real-world uses for an effective time series model for predicting tax revenue:

a. Government budgeting: An effective tax revenue forecasting model can assist governments in making financial decisions, planning expenditures, and efficiently allocating resources.

b. Tax planning and policy-making: Forecasting tax income can help policymakers make informed judgments about future policy and get insight into the efficiency of current tax laws.

c. Economic forecasting: Because of the strong relationship between tax revenue and economic activity, an effective forecasting model can be utilized to predict changes in the economy and guide the decisions that businesses make.

d. Resource allocation: Based on projected revenue, resources can be distributed to a variety of areas, including education, healthcare, and public safety.

e. Public debt management: Forecasting tax revenue accurately is crucial for managing the public debt well. It enables decision-makers to choose the right degree of borrowing and debt repayment.

Overall, an efficient time series model for tax revenue forecasting has many practical applications in real life, from government budgeting to economic forecasting and public debt management.

REFERENCES

Aamir, M., Qayyum, A., Nasir, A., Hussain, S., Khan, K. I., & Butt, S. (2011). Determinants of Tax Revenue: A Comparative Study of Direct taxes and Indirect taxes of Pakistan and India. *International Journal of Business and Social Science* , 1–4.

Ajayi, G. O., Adesina, A. A., & Adewuyi, A. A. (2013). Forecasting crude oil revenue in Nigeria. *Journal of Applied Sciences and Environmental Management*, 67–72.

Akhter, T., & Hassan, Md. H. (2012). Public Debt Burden and Economic Growth of Bangladesh: A VAR Approach. *SSRN Electronic Journal*, *9*(4), 5–12. https://doi.org/10.2139/ssrn.2152592

Alashari, M., El-Rayes, K., Attalla, M., & Al-Ghzawi, M. (2022). Multivariate time series and regression models for forecasting annual maintenance costs of EPDM roofing systems. *Journal of Building Engineering*, *54*(1), 4–10. https://doi.org/10.1016/j.jobe.2022.104618

Bala, D. A., & Shuaibu, M. (2022). Forecasting United Kingdom's energy consumption using machine learning and hybrid approaches. *Energy and Environment*, *1*(1), 10–23. https://doi.org/10.1177/0958305X221140569/ASSET/IMAGES/LARGE/10.1177_0958305X221140569-FIG14.JPEG

Chaudhry, I., & Munir, F. (2010). Determinants of Low Tax Revenue in Pakistan. *Pakistan Journal of Social Sciences*, *30*(2), 2–13.

Chen, W. T., & Chen, C. F. (2017). Forecasting Taiwan's tax revenues using ARIMA models. *International Journal of Economics, Commerce and Management*, *V*(2), 5–15.

Erdoğdu, H., & Yorulmaz, R. (2019). Comparison of Tax Revenue Forecasting Models for Turkey. In *34. International Public Finance Conference* (Vol. 1, pp. 2–9). Istanbul University Press. https://doi.org/10.26650/PB/SS10.2019.001.075

Eugene, N., & Abigail, E. C. (2016). Effect of Tax Policy on Economic Growth in Nigria (1994-2013). *International Journal of Business Administration*, *7*(1), 1–8. https://doi.org/10.5430/ijba.v7n1p50

FIRS. (2022). *What is tax ?* Https://Www.Firs.Gov.Ng/.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (Vol. 2). Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Himani. (2016). Determinants of tax revenue in India. *International Journal of Research in Economics and Social Sciences*, *1*(1), 2–12.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Jabeur, S. Ben, Ballouk, H., Mefteh-Wali, S., & Omri, A. (2022). Forecasting the macrolevel determinants of entrepreneurial opportunities using artificial intelligence models. *Technological Forecasting and Social Change*, *175*(121353), 8–10. https://doi.org/10.1016/j.techfore.2021.121353

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd ed., Vol. 2). Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Jang, S.-B. (2019). A Design of a Tax Prediction System based on Artificial Neural Network. *2019 International Conference on Platform Technology and Service (PlatCon)*, 1–4. https://doi.org/10.1109/PlatCon.2019.8669416

Karagöz, K. (2013). Determinants of tax revenue: does sectorial composition matter? *Journal of Finance, Accounting and Management*, *4*(1), 50–63.

Luqman, O. S. (2014). The Impact of Value Added Tax on Revenue Generation in Nigeria. *SSRN Electronic Journal*, *1*(1), 1–8. https://doi.org/10.2139/ssrn.2513207

Martinez-Vazquez, & McNab, R. M. (2010). Tax revenue and economic growth in Latin America. *Journal of Development Studies*, *1*(1), 2–8.

Matta, C. E. da, Bianchesi, N. M. P., Oliveira, M. S. de, Balestrassi, P. P., & Leal, F. (2021). A comparative study of forecasting methods using real-life econometric series data. *Production*, *31*, 01–15. https://doi.org/10.1590/0103-6513.20210043

Micheni Nelson Kirimi, Atitwa Edwin Benson, & Kimani Patrick. (2022). Forecasting Domestic Tax Revenues in Kenya Using Sarima & Holt-Winters Methods. *International Journal of Information Management Sciences*, *6*(2022), 3–10.

Otu, A., George A., O., Jude, O., Hope Ifeyinwa, M., & Andrew I., I. (2014). Application of Sarima Models in Modelling and Forecasting Nigeria's Inflation Rates. *American Journal of Applied Mathematics and Statistics*, *2*(1), 16–28. https://doi.org/10.12691/ajams-2-1-4

Petrovski, A., Petruseva, S., & Zileska Pancovska, V. (2015). Multiple Linear regression model for predicting bidding price. *Technics Technologies Education Management*, *10*(1), 386–393.

Qadirpatoli, A., Zarif, T., & Syed, N. (2012). Impact of Inflation on Taxes in Pakistan: An empirical study of 2000-2010 period Jel classification : E310. *IBT Journal of Business Studies*, *8*(1), 1–8.

Rhanoui, M., Yousfi, S., Mikram, M., & Merizak, H. (2019). Forecasting Financial Budget Time Series: ARIMA Random Walk vs LSTM Neural Network. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *8*(4), 1–8. https://doi.org/10.11591/ijai.v8.i4.pp317-327

Rob J Hyndman, & George Athanasopoulos. (n.d.). *Forecasting: Principles and Practice (2nd ed)*. Retrieved September 6, 2022, from https://otexts.com/fpp2/

Sepp Hochreiter, & J□urgen Schmidhuber. (1997). *LONG SHORT-TERM MEMORY*. Neural Computation. https://www.bioinf.jku.at/publications/older/2604.pdf

Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications* (Vol. 4). Springer International Publishing. https://doi.org/10.1007/978-3-319-52452-8

Siami-Namini, S., Muhammad, D., & Fahimullah, F. (2018). The Short and Long Run Effects of Selected Variables on Tax Revenue - A Case Study. *Applied Economics and Finance*, *5*(5), 23. https://doi.org/10.11114/aef.v5i5.3507

Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2019). A Comparison of ARIMA and LSTM in Forecasting Time Series. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, 1394–1401. https://doi.org/10.1109/ICMLA.2018.00227

Streimikiene, D., Raheem Ahmed, R., Vveinhardt, J., Ghauri, S. P., & Zahid, S. (2018). Forecasting tax revenues using time series techniques – a case of Pakistan. *Economic Research-Ekonomska Istraživanja*, *31*(1), 4–20. https://doi.org/10.1080/1331677X.2018.1442236

Valerie Watts. (2022). *Coefficient of Multiple Determination – Introduction to Statistics*. https://ecampusontario.pressbooks.pub/introstats/chapter/13-4-coefficient-of-multiple-determination/

WALTER ENDERS. (2015). *Applied econometric time series* (Vol. 4). John Wiley & Sons.

Wilson, G. T. (2016). Time Series Analysis: Forecasting and Control, 5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1. *Journal of Time Series Analysis*, *37*(5), 709–711. https://doi.org/10.1111/JTSA.12194

Yaser S. Abu-Mostafa, Malik Magdon-Ismail, & Hsuan-Tien Lin. (2012). Learning From Data. In *Department of Electrical Engineering, California Institute of Technology* (Vol. 1).

Zhang, R., Song, H., Chen, Q., Wang, Y., Wang, S., & Li, Y. (2022). Comparison of ARIMA and LSTM for prediction of hemorrhagic fever at different time scales in China. *PLOS ONE*, *17*(1), 3–9. https://doi.org/10.1371/journal.pone.0262009

Zhang, X., Xue, T., & Eugene Stanley, H. (2019). Comparison of Econometric Models and Artificial Neural Networks Algorithms for the Prediction of Baltic Dry Index. *IEEE Access*, *7*, 2–11. https://doi.org/10.1109/ACCESS.2018.2884877

APPENDIX

Table 1: Full Data set used for this research in Billion Naira

Year	TOTALREV	PPT	CIT	VAT	CGT	EDT	SD	CONS	NITDEF
2010-01	225.0739	111.0819	50.8026	48.0979	0.028	11.1058	0.4416	2.5188	0.9973
2010-02	173.3886	83.8694	39.5541	46.4374	0.1174	0.6015	0.3655	2.3717	0.0716
2010-03	173.8879	73.5519	43.1208	51.1445	0.0676	2.3402	0.6532	2.9646	0.0451
2010-04	212.2371	109.3688	46.0926	49.755	0.0346	4.0036	0.4394	2.45	0.0931
2010-05	198.2065	112.5318	39.4148	41.5267	0.2816	1.6328	0.3839	2.4292	0.0057
2010-06	265.7049	124.2117	67.9681	54.1947	0.0628	11.9663	0.744	3.6477	2.9096
2010-07	226.4147	109.7281	63.1902	44.6255	0.1258	4.4803	0.5778	2.6923	0.9947
2010-08	265.7115	120.9611	79.8292	48.5183	0.1534	13.1974	0.4966	2.4772	0.0783
2010-09	264.3625	129.0012	60.5617	42.7534	0.0925	28.7837	1.0797	2.0702	0.0201
2010-10	222.64	120.5255	50.9345	45.6765	0.0599	1.7393	0.3821	2.726	0.5962
2010-11	314.0708	188.4935	68.9244	44.7018	0.0124	8.2739	0.5791	3.0407	0.045

2010-12	297.686	197.039	48.1096	47.4599	0.0007	1.0532	0.4505	3.5437	0.0294
2011-01	336.8067	236.7522	41.5287	50.7931	0.1997	2.8029	0.4164	3.8545	0.4592
2011-02	342.4236	254.7853	34.1034	49.1554	0.825	0.7018	0.4097	2.4346	0.0084
2011-03	275.961	165.8922	38.911	61.6073	0.2572	3.3752	0.5991	5.2097	0.1093
2011-04	290.0848	184.1902	50.6491	44.3371	4.5706	2.715	0.4552	3.1474	0.0202
2011-05	256.0543	156.9412	37.4567	53.7482	0.2425	3.9212	0.562	3.1343	0.0482
2011-06	439.1668	307.0558	64.3151	54.6924	0.2861	7.2266	0.5785	3.5709	1.4414
2011-07	491.4071	306.7827	84.2335	65.9933	0.0476	25.6495	0.4462	3.1894	5.0649
2011-08	501.3514	315.7837	103.882	57.0065	2.7484	16.2901	0.4189	4.7249	0.4969
2011-09	424.1665	265.6614	78.5055	60.7393	0.0181	14.6581	0.6275	3.5758	0.3808
2011-10	412.1296	277.1868	67.978	51.6384	0.0043	11.0476	0.622	3.3483	0.3042
2011-11	390.5704	269.753	51.8273	52.3145	0.025	12.2409	0.6582	3.4617	0.2898
2011-12	468.3535	329.8068	46.285	57.1281	0.08	30.113	0.6686	4.2202	0.0518
2012-01	369.388	267.3912	41.1845	53.2406	0.0307	3.8534	0.3644	3.3191	0.0041

2012-02	466.8679	365.2414	35.2779	60.608	0.086	1.761	0.4335	3.1578	0.3023
2012-03	336.2123	216.1164	48.4564	62.0089	0.4711	3.0074	0.59	4.765	0.7971
2012-04	385.171	265.5589	44.6283	59.6784	1.6993	7.685	0.4815	3.9126	1.527
2012-05	328.0206	198.797	59.5908	60.0843	0.7839	2.2873	0.7117	4.876	0.8896
2012-06	554.0379	277.9597	186.4961	59.2196	0.2862	15.8549	0.679	8.5334	5.009
2012-07	524.1224	301.5009	134.5455	53.9105	4.0975	24.8104	0.983	4.0414	0.2332
2012-08	438.0871	245.1543	70.9682	60.4372	0.028	57.1335	0.5698	3.7467	0.0494
2012-09	403.8829	249.8554	50.538	56.3425	0.0347	42.9915	0.5296	3.447	0.1442
2012-10	454.8635	289.7126	71.141	64.7667	0.7345	23.7037	0.7867	3.8568	0.1615
2012-11	402.571	290.2571	43.484	62.7226	0.1396	1.3529	0.7035	3.901	0.0103
2012-12	344.4282	233.7746	43.9818	57.5358	0.5251	3.9945	0.5501	4.0561	0.0102
2013-01	478.9602	330.8139	61.6704	65.2909	0.0523	16.8867	0.5133	3.6909	0.0418
2013-02	378.0608	249.6329	50.3976	62.7068	0.0991	10.2613	0.5413	4.3775	0.0443
2013-03	349.7104	220.2028	44.5173	64.1987	0.0153	15.7525	1.0266	3.9083	0.0889

2013-04	357.6869	247.2323	47.4036	54.5713	2.0478	1.5616	0.6059	3.8203	0.4441
2013-05	431.2817	258.1387	86.2623	74.8728	0.6702	3.3443	0.7551	4.0657	3.1726
2013-06	647.5374	288.0547	269.1176	51.1703	14.0654	15.8459	0.4601	3.8099	5.0135
2013-07	397.5788	142.9701	102.3032	74.1974	0.0659	72.7899	0.6499	3.9739	0.6285
2013-08	421.0056	188.3881	87.9013	68.9323	0.0515	71.6839	0.4644	3.5149	0.0692
2013-09	337.1439	189.1229	51.5485	63.941	0.0221	28.3942	0.431	3.6425	0.0417
2013-10	361.1241	200.4935	49.1091	66.3462	2.4887	37.9983	0.8268	3.6834	0.1781
2013-11	350.1665	187.1516	64.0978	91.7303	0.0326	2.3669	0.5178	4.1951	0.0744
2013-12	295.3857	164.1654	56.849	64.7255	0.045	2.4732	0.8103	6.2575	0.0598
2014-01	346.5017	191.0479	60.219	82.2767	0.0067	7.6725	1.3673	3.8651	0.0465
2014-02	401.3725	268.2075	58.066	66.8012	0.0008	2.4162	0.7509	5.0798	0.0501
2014-03	308.4999	178.8329	58.3090552	63.3075	0.7763	2.5526	0.6989	3.9753	0.0473
2014-04	312.4967	177.2038	61.4716	65.4257	0.1128	2.0475	0.5271	4.4481	1.2601
2014-05	569.215	215.6306	281.5703049	65.4154	0.0117	1.4065	0.7215	3.9378	0.5212

2014-06	573.4587	246.4339	215.7401	66.414	0.1659	34.6086	0.6199	3.6483	5.828
2014-07	410.3682	207.7747	93.9348	65.4674	0.2094	34.9272	0.9151	5.7144	1.4252
2014-08	408.9691	193.3894	109.0666	61.5129	1.1003	38.7545	0.8393	3.9771	0.329
2014-09	379.8957	193.6354	74.6853	65.1022	0.2094	41.4919	0.6147	3.9341	0.2227
2014-10	346.2972	194.1511	61.1501	67.1373	0.021	18.767	0.7043	4.2846	0.0818
2014-11	339.3691	206.9194	63.7799	60.6386	0.016	2.0187	1.7976	4.1678	0.0311
2014-12	318.1166	180.7208	53.2477	73.4658	0.0195	2.9505	1.387	6.2601	0.0652
2015-01	291.9092	147.3773	69.9335	63.9354	0.1514	3.6028	0.4896	6.3332	0.086
2015-02	272.9741	150.4805	56.7156	58.2566	0.0455	1.0611	0.915	5.4468	0.053
2015-03	217.4847	93.1792	43.5237	71.1973	0.0533	2.5178	0.5836	6.4003	0.0295
2015-04	195.7858	50.6705	62.1301	75.1603	0.0032	2.015	0.4142	4.2711	1.1214
2015-05	246.1236	118.9016	57.7133	56.8212	1.7246	4.2336	0.6361	4.7528	1.3404
2015-06	746.2154	136.5696	501.6561	64.9922	10.2796	20.5973	0.5164	4.7471	6.8571
2015-07	408.0853	103.0329	155.7467	74.9451	3.9552	63.0913	0.8231	5.0545	1.4365

2015-08	322.147	110.8738	104.7933	62.1765	0.0263	38.8001	0.4935	4.8667	0.1168
2015-09	250.2506	111.9589	66.1135	56.399	0.2634	10.2561	0.7662	4.4093	0.0842
2015-10	338.5196	70.1293	162.8366	60.1973	0.1253	41.2554	0.4588	3.4429	0.074
2015-11	232.7114	103.0385	52.8699	61.1813	0.1615	11.7257	0.3833	3.3372	0.014
2015-12	219.5507	93.7486	50.5141	62.0713	0.0127	6.884	0.6047	4.6795	1.0358
2016-01	201.0415	52.3516	71.4687	69.7192	0.0161	3.6202	0.3632	3.4289	0.0736
2016-02	200.3464	73.9884	53.8827	64.7811	0.0216	2.3562	0.6527	4.6396	0.0241
2016-03	162.4818	50.4078	41.5052	64.234	0.1903	2.2656	0.4007	3.4468	0.0314
2016-04	168.7642	38.2414	55.8151	65.2593	0.0458	3.0664	0.4324	4.6229	1.2809
2016-05	211.868	51.0454	69.713	65.1163	11.9624	7.207	0.4237	4.8743	1.5259
2016-06	614.2843	238.8048	222.4562	67.4009	60.5849	16.3025	0.4284	5.1325	3.1741
2016-07	343.9712	94.141	126.7696	66.9871	24.1481	27.0381	0.486	4.1053	0.296
2016-08	367.1517	122.5738	116.1125	75.9621	0.0177	46.1043	0.4423	5.753	0.186
2016-09	281.164	106.8642	96.9051	64.2648	0.023	7.3223	0.4762	5.2729	0.0355

2016-10	252.7202	115.9809	49.4515	69.6212	2.0746	9.297	0.3931	5.8224	0.0795
2016-11	253.6207	114.2965	55.2922	75.5795	0.2575	1.5049	0.7818	5.8892	0.0191
2016-12	250.0474	99.1123	60.0436	79.2736	0.0614	4.0382	0.6225	6.8739	0.0219
2017-01	317.8386	142.3301	61.4744	73.5216	0.0838	31.3138	0.6237	8.4887	0.0025
2017-02	229.5264	111.0827	40.2554	69.2076	0.0014	1.0696	0.4854	7.3628	0.0615
2017-03	230.8285	84.8862	54.2287	78.6513	0.0254	1.5469	1.5242	9.8506	0.1152
2017-04	218.3565	65.5978	55.8984	84.6736	0.0141	1.8187	0.6496	9.3613	0.343
2017-05	286.7972	84.8223	103.5914	79.985	0.019	4.2222	0.6098	10.5011	3.0464
2017-06	499.022	147.4514	236.0518	81.6447	0.7927	18.8972	0.5615	8.678	4.9447
2017-07	329.7519	84.1971	120.8101	80.5335	1.78	32.5516	0.455	8.0729	1.3517
2017-08	469.7431	192.5567	167.347	86.7122	0.0443	11.4627	0.8223	10.7362	0.0617
2017-09	315.8337	113.9507	96.7774	83.315	0.0206	16.0862	0.7226	4.9247	0.0365
2017-10	332.3291	138.6823	65.9029	89.7135	0.1703	30.598	0.8856	6.2868	0.0897
2017-11	390.9574	166.5534	132.4634	80.4266	0.0587	1.9213	0.6623	8.842	0.0297

2017-12	406.9608	188.371	115.0945	83.9638	0.17	3.4692	0.9315	14.9094	0.0514
2018-01	384.9874	180.8981	74.927	96.6465	0.0309	20.7274	0.7768	10.8829	0.0978
2018-02	459.0512	296.0873	63.1301	89.447	0.0608	1.6949	1.7366	6.8616	0.0329
2018-03	329.5753	167.7897	65.63	83.7003	0.2225	3.3751	1.7446	7.0809	0.0322
2018-04	440.4434	197.5911	141.933	87.9658	0.1673	1.6533	0.7839	10.2731	0.0759
2018-05	406.9408	186.7614	103.7778	93.4233	0.031	9.5907	0.8514	8.682	3.8232
2018-06	486.8602	139.4998	225.8214	85.3426	5.968	18.4527	0.9462	5.642	5.1875
2018-07	477.618	160.6269	159.3947	82.3133	5.7473	59.4265	0.8942	6.9447	2.2703
2018-08	524.9513	265.9858	110.5112	114.5423	0.0777	27.1228	1.454	5.136	0.1215
2018-09	377.8482	199.7712	92.7604	76.6484	0.0185	3.2475	1.2823	4.0728	0.0471
2018-10	533.4811	276.0552	90.594	105.1717	0.0476	52.6155	1.84	7.1182	0.0389
2018-11	493.6284	232.5374	159.0885	92.0788	0.0948	2.8868	1.712	5.1912	0.0389
2018-12	405.5062	163.9768	128.7491	100.76	0.1283	2.4916	1.7754	7.538	0.0871
2019-01	364.4139	156.8787	93.3076	104.4687	0.0405	2.3833	1.2447	6.0683	0.0221

2019-02	359.6581	182.9869	70.9317	96.3892	0.0152	1.8196	1.3438	6.1389	0.0328
2019-03	322.8178	153.3543	68.566	92.1815	0.0407	3.0267	0.7982	4.7942	0.0562
2019-04	387.9838	197.666	80.6638	96.4856	0.1744	4.3146	1.0937	5.475	2.1107
2019-05	403.208	99.1019	179.7068	106.8265	0.2504	4.6606	1.6679	8.9094	2.0845
2019-06	609.4168	206.2256	250.5607	108.6309	0.5504	29.4367	0.9571	4.1018	8.9536
2019-07	542.9361	187.2624	194.32	94.1595	0.7749	58.5617	1.3917	5.2222	1.2437
2019-08	535.6928	211.9344	202.4196	88.0824	0.1565	25.3099	1.1449	6.4102	0.2349
2019-09	485.9398	193.3507	124.1462	92.8742	0.3672	69.9772	1.1603	3.8815	0.1825
2019-10	455.2913	191.7401	135.01	104.91	0.5606	14.6956	1.7905	6.4901	0.0944
2019-11	345.2586	151.5783	90.0094	90.1666	0.3449	4.0712	2.923	6.1095	0.0557
2019-12	449.2993	182.1891	136.9869	114.806	2.7013	2.8006	2.6762	7.0291	0.1101
2020-01	371.1579	171.1632	84.545	104.7584	0.1659	5.1219	1.4599	3.9014	0.0422
2020-02	328.96	156.6545	62.7412	99.5521	0.1764	2.8168	1.7175	5.2779	0.0236
2020-03	474.9341	194.5021	148.3816	120.2686	0.301	3.4008	1.5734	5.8812	0.6254

2020-04	381.4083	163.6865	55.879	94.496	0.0073	2.6685	59.885	2.6289	2.1572
2020-05	380.3635	182.7799	85.7351	103.873	0.0266	2.3003	0.8641	4.3422	0.4423
2020-06	526.5905	93.835	260.4199	128.8264	0.5835	27.5151	1.8339	4.6259	8.9508
2020-07	530.3847	192.7011	133.472	132.6195	0.4153	60.592	2.0973	5.6606	2.8269
2020-08	490.4744	52.4938	197.1128	150.2301	1.2909	77.7718	3.0339	5.9846	2.5565
2020-09	399.1348	107.9176	85.4242	141.8585	0.0775	56.7368	2.1261	4.8346	0.1595
2020-10	332.3633	43.7337	100.1825	126.4633	0.0651	12.0621	39.7728	10.0163	0.0675
2020-11	356.1353	93.1263	93.6147	156.8671	0.1329	3.4422	2.1344	6.7001	0.1176
2020-12	380.3369	64.3855	101.9248	171.3579	0.2762	5.1351	3.6587	33.5169	0.0818
2021-01	396.6774	113.7465	98.3464	157.3513	0.0716	2.5435	1.9861	5.5855	0.0067
2021-02	371.6062	137.5825	56.4841	157.3267	0.5511	2.1467	3.0458	8.1655	0.0292
2021-03	517.579	75.9044	237.8207	181.712	0.1245	6.7142	2.5843	8.9166	0.6255
2021-04	413.5321	100.3984	122.5914	176.7099	0.107	5.9251	2.2414	3.7548	0.1493
2021-05	366.0905	57.1116	117.5993	181.0779	0.5691	3.4981	2.042	2.4662	0.3594

2021-06	696.979	159.4019	216.7975	154.4651	0.4596	19.7383	0.2071	9.4685	3.4681
2021-07	427.6895	64.865	169.2297	151.1344	0.6045	26.5032	0.407	2.9288	0.3099
2021-08	444.0541	56.5082	155.4938	178.5092	0.1139	37.2813	0.0512	2.8629	0.0572
2021-09	558.6378	183.764	147.7997	170.8501	0.1388	28.7218	0.0251	2.6284	10.7226
2021-10	340.4708	53.8636	96.5286	166.2845	0.0557	3.9026	7.2821	2.582	0.1665
2021-11	502.7552	170.982	113.345	196.1754	0.2338	3.7342	4.4283	2.689	0.1356
2021-12	532.1541	103.373	137.936	201.2554	1.1593	48.8262	9.6399	9.1498	3.277