iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VIII **Month of publication:** Aug 2023

DOI: https://doi.org/10.22214/ijraset.2023.55423

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VIII Aug 2023- Available at www.ijraset.com

Business Data Revenue Analysis

Alesha Bijoy[1], Bhagya Babu[2], P V Mahesh Anand[3]
[1, 2, 3]*Department of Mathematics, Amrita Vishwa Vidyapeetham, Kochi, Kerala, India*

Abstract: ***Forecasting time series data is an important subject in economics, business, and finance. This research paper aims to explore the effectiveness of three different time series forecasting techniques, namely ARIMA, SARIMA, and LSTM, along with a machine learning algorithm, Random Forest, in predicting revenue for a business organization. The study uses historical revenue data from a selected company and evaluates the performance of each model based on statistical metrics like root mean squared error (RMSE). The paper provides comparison of the accuracy of the four techniques. Root mean square error (RMSE) of LSTM was 42, 268.2197, Random Forest was 19398.4725 while ARIMA and SARIMA had 111413.5821 and 338174.3004 respectively. The findings of the study reveal that LSTM outperforms the traditional ARIMA and SARIMA models, but in terms of accuracy in revenue forecasting Random Forest outperforms all. However, the study also highlights the importance of selecting the appropriate model based on the nature and complexity of the data, as well as the availability of computational resources. The findings of this research can assist businesses in making informed decisions about revenue forecasting methods and contribute to the development of more effective revenue forecasting models.***

I. INTRODUCTION

Forecasting revenue is a critical aspect of any business as it helps organizations in strategic decision making, budgeting, and planning. In the context of an audit firm, forecasting revenue can be particularly challenging due to the nature of their services, which are heavily influenced by external factors such as changes in regulations, economic conditions, and market demand. Time series forecasting techniques have been widely used to forecast revenue in various industries, including audit firms. In this research paper, we aim to investigate the effectiveness of several time series forecasting techniques, namely ARIMA, SARIMA, LSTM, and Random Forest, for forecasting revenue time series data of an audit firm.

II. LITERATURE REVIEW

A key component of corporate planning and decision-making is revenue forecasting. The efficiency of various forecasting strategies in projecting income has been examined in a number of studies[Cha02][Tsa10]. We will look at several pertinent papers on ARIMA, SARIMA, LSTM, and Random Forest revenue forecasting in this literature study.

The effectiveness of three forecasting techniques for predicting the inventory levels of power plant spare parts in the short term is compared in the article ”Comparison of ARIMA, Linear Trend, and Single Exponential Smoothing for short-term forecasting of power plant spare parts inventory” by Utomo, A., Hasbullah, & Hasibuan (2017)[SHH17].From models they considered like ARIMA, Linear Trend and Single exponential Smoothing, ARIMA performs best in accuracy measures, demonstrating its suitability for short-term forecasting. The application of Random Forest Regression (RFR) for short-term power demand forecasting is examined in the study ”Short-term electricity load forecasting using random forest regression” by Liu, J., Zhu, P., & Zhao,J(2018)[LZZ18].In order to anticipate electricity load, this study compares the performance of RFR to other machine learning techniques including Support Vector Regression (SVR) and Artificial Neural Networks (ANNs). The authors also look at how various input variables affect forecasting accuracy. The sensitivity analysis revealed that the input factors that had the greatest impact on forecasting ability were the temperature and the day of the week. The findings highlight the significance of input variables in achieving accurate and trustworthy forecasting results and show the superiority of RFR over SVR and ANNs for short-term power load forecasting. The effectiveness of two well-known time series forecasting techniques, ARIMA and Long Short-Term Memory (LSTM) neural networks, was compared in the study ”Forecasting Economics and Financial Time Series: ARIMA vs. LSTM” by Namin, S. S., & Namin, A. S. (2018)[NN18]. The authors conducted a thorough literature assessment of studies that applied these techniques for fore-casting various economic and financial time series, including stock prices, exchange rates, and GDP, in order to compare the effectiveness of ARIMA and LSTM. Various statistical indicators, such as MAE, RMSE, and MAPE, demonstrate forecasting accuracy, with different measures being appropriate for different contexts. The author of the paper ”The Application of Forecasting Sales of Services to Increase Business Competitiveness” by Kolkova, A. (2018)[Kol18] did a review of the literature on studies that employed forecasting techniques to predict service sales to boost business competitiveness.

The author emphasizes the use of proper performance indicators for accurate sales projections. Forecasting techniques must be chosen considering the specific business requirements and features of the service sector.

The usage of random forest models in predicting in the Czech Republic was reviewed in the article ”Random forest as a model for Czech forecasting” by Gawthorpe, K. (2019)[Gaw19].According to the study, random forest models have been applied to a variety of tasks, including forecasting weather, energy demand, and the state of the economy. The primary benefit of random forest models is their propensity to accurately handle huge datasets without overfitting. The usage of machine learning algorithms for rainfall prediction in various parts of the world was reviewed in the paper ”Comparative study of machine learning algorithms for rainfall prediction - A case study in Nepal” by Paudel, N., and Yogi, T. N. (2019)[PY19]. The study demonstrated that the Random Forest method beat other algorithms in terms of statis-tical metrics including root mean square error, mean absolute error, and coefficient of determination. Other algorithms compared in the study were Artificial Neural Network and Support Vector Machine. The study came to the conclusion that Random Forest might be a useful tool for predicting rainfall in Nepal and other places with comparable features. The research ”Forecasting hotel daily room demand with transformed data using time series methods” by Phumchusri and Suwatanapongched (2019)[PS19] discusses a number of time series techniques that have been applied to forecast hotel room demand, including moving average (MA), exponential smoothing (ES), and autoregressive integrated moving average (ARIMA) models. The authors also go over the value of modifying the data to take seasonality and patterns into consideration as well as the application of machine learning methods like artificial neural networks (ANN) and support vector regression (SVR) to forecasting hotel room demand. The SARIMA model with Box-Cox transformation emerges as a suitable strategy.

III. METHODOLOGY

The methodology encompasses data acquisition, data preparation, data cleaning, training and testing of models, as well as forecasting outputs. We used four algorithms in forecasting revenue of an organisation which includes three time series forecasting algorithms : ARIMA, SARIMA ,and LSTM, along with with a machine learning algorithm, Random Forest. All of these models are examined to find which one is better at forecasting . Excel and Python were the tools we used in this study.

A. Data Acquisition

The primary data obtained is dated from August 2019 to March 2022. The data was given three separate folders with details of each financial year and each month as excel files. Each file contained Journal Bills of the month with client information and the corresponding amount credited.

B. Data Preparation

Collected data was manually entered into an Excel sheet, documenting services offered by the company. This process involved frequent visits for accurate metadata. The dataset has 248 rows, 31 columns, with 24 columns denoting service verticals. Jupyter Notebook and Pandas were used for analysis (code: pd.read_excel('data.xlsx')). Verticals were tagged with 1 for service identification.

C. Data Cleaning

As the data originated from SVJS Associates & Company Secretaries' affiliated entity, revenue records were accessible only through journal bills. Thus, reconciling revenue and monthly statements demanded manual effort. Spell errors and service overlaps in journal bills required rectification after company verification. Tax and legal form-related fees, marked in red, weren't considered income. Yellow-marked rows required clarification. Due to sensitivity, client names were coded, with extra spaces eliminated for Python. Omitting 'Professional Category,' two datasets were prepared: data1 for Random Forest and LSTM, Data2 for ARIMA and SARIMA. Data2 was monthly aggregated, sustaining continuity. Empty rows were zero-filled. An allocation of 10000 from Feb 2021 countered tax-related removal, aligning with revenue statements.

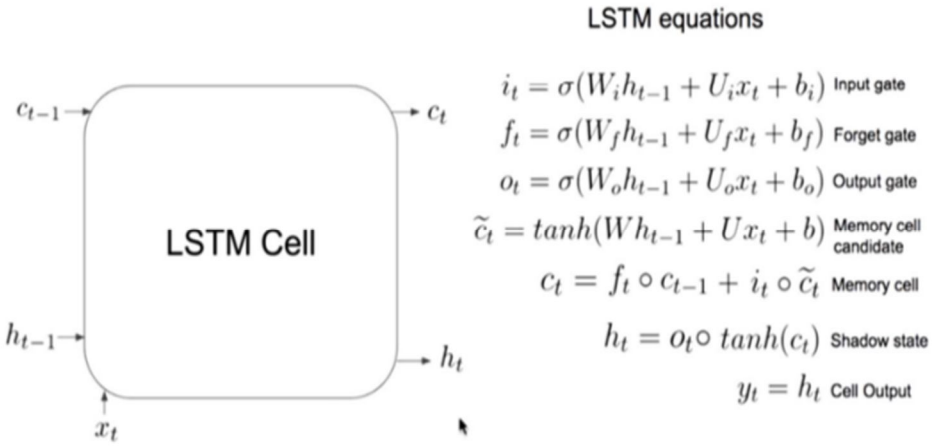
D. Algorithms Used

1) *ARIMA* : The widely used Autoregressive Integrated Moving Average (ARIMA) model is employed for time series prediction. Comprising autoregression (AR), differencing (I), and moving average (MA) components, ARIMA captures current value dependence on past values, models error terms linearly, and employs differencing to eliminate trends or seasonality. ARIMA is effective for stationary time series data. The complete model is represented as:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

Here, ϵ signifies the model's error term. The ARIMA(p,d,q) model integrates these three components, resulting in a time series model incorporating autoregressive and moving average aspects, alongside differencing for stationarity. denotes the dependent variable at any time, involving lagged observations, represents lag 1 of the series, and β terms originate from the AR component, while ϵ pertains to error terms within the MA model.

2) *Sarima :* The Seasonal Autoregressive Integrated Moving Average (SARIMA) model extends ARIMA to manage time series seasonality. It incorporates seasonal elements like seasonal autoregression (SAR) and seasonal moving average (SMA) terms. The SARIMA model is defined as SARIMA(p, d, q)(P, D, Q)m, with P, D, Q, and m representing additional seasonal components, where m is seasons per year. SARIMA suits data displaying seasonal patterns.
3) *LSTM:* Long Short-Term Memory (LSTM) stands as a recurrent neural network, adept at modelling intricate time series data. Equipped with a memory cell capable of retaining information for extended periods, the LSTM effectively captures prolonged dependencies within the data. Incorporated gating mechanisms—input, forget, and output gates—manage information flow to and from the memory cell. The LSTM excels with non-linear, non-stationary time series. The LSTM cell structure and gate equations are outlined below.



1) *Random Forest:* Random Forest, a versatile machine learning algorithm, excels in classification and regression tasks. Comprising decision tree ensembles, each tree is trained on random data and feature subsets. Model output is the mean of individual tree outputs. It's particularly effective for non-linear, non-stationary time series data.

E. Exploratory Data Analysis

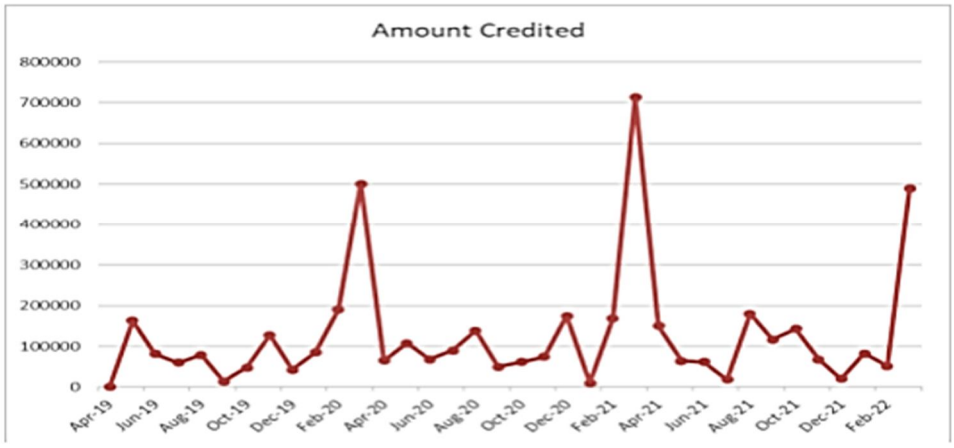


Figure 1 : Monthly summation of revenue from April 2019 to March 2022

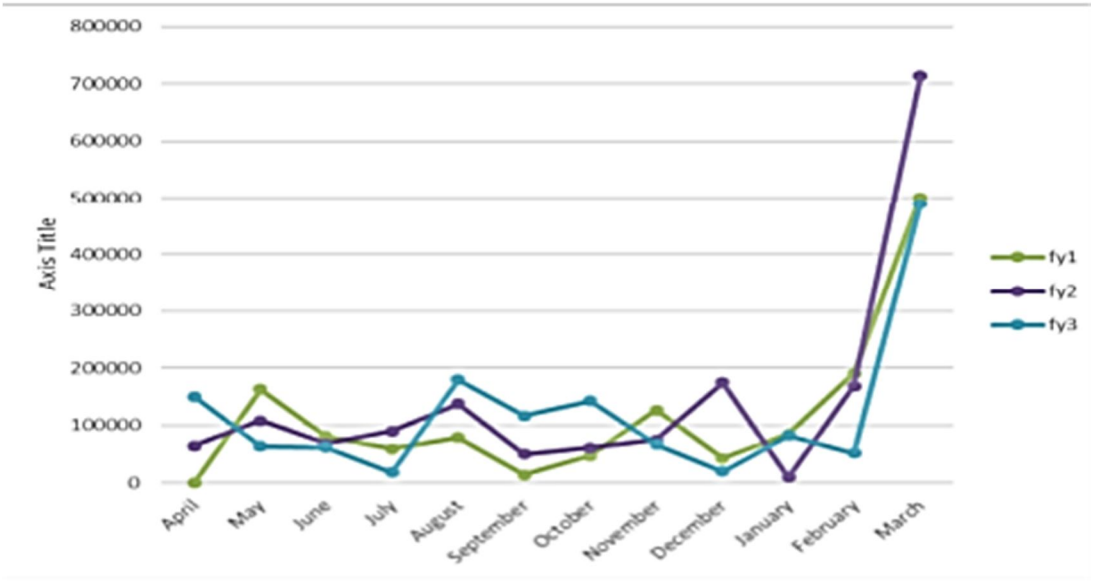


Figure 2 : Data fluctuation in a financial year over the three years

IV. RESULT ANALYSIS

In this study, the performance of each model was evaluated using a comprehensive RMSE and MAE metrics to assess their accuracy and predictive power. The following is a table of the result analysis for each model:

Model	MAE	RMSE
Random Forest	6282.8404	19398.4725
LSTM	31289.8078	42268.3197
ARIMA	241111.20	111413.58
SARIMA	79499.98	338174.3

V. CONCLUSION

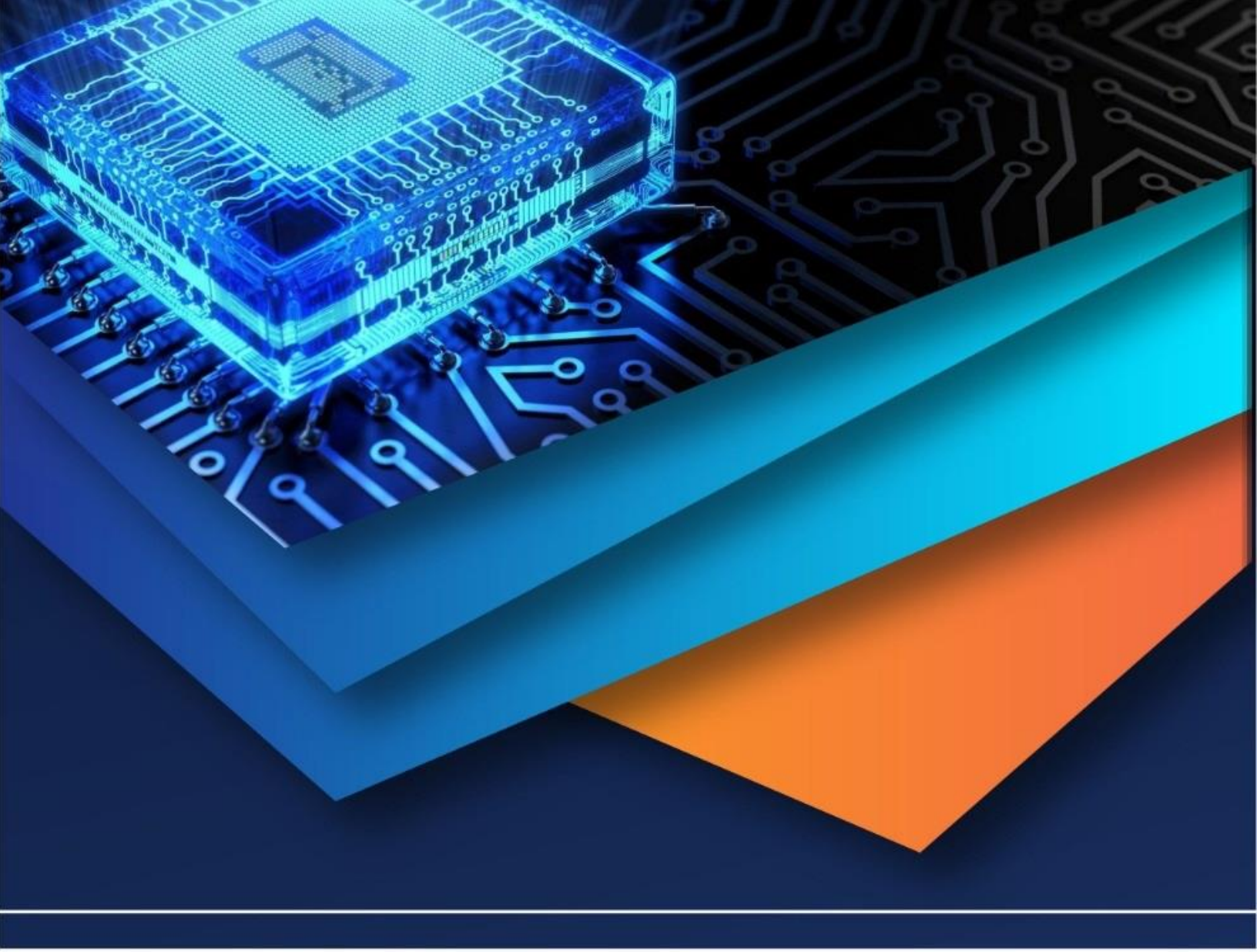
The results showed that Random Forest was superior to LSTM followed by ARIMA and SARIMA when these techniques were used on a set of financial data. More specifically, the prediction accuracy of the models are given as, SARIMA has RMSE value 338174.3, ARIMA has RMSE value 111413.58 while Random Forest and LSTM has RMSE values 19398.4725 and 42269.2197 respectively.

The average absolute difference between the expected and actual values is measured by MAE. A model is better for predicting revenue if it has a lower MAE value, which signifies that it makes less errors in its predictions. In this study Random Forest had minimum MAE value of 6282.8404 followed by LSTM with 31289.8078 , SARIMA with 79499.98 and ARIMA with highest of 241111.20.Therefore according to this metric it is found that Random Forest made less errors among the models compared, which in turn makes Random Forest, an ensemble model is the preferable model for revenue forecasting.

The dataset should contain historical revenue data as well as other pertinent factors that may have an impact on revenue in order to do revenue forecasting using LSTM and Random Forest. Thus, in order to allow the models to accurately capture long-term trends and seasonal patterns, the dataset should have a sufficient number of data points over a sizable time period. Additionally, there should be no errors, outliers, or missing values in the dataset. To avoid model biases, it is also preferable to have a balanced dataset and to properly assess the performance of the models, it is better to divide the dataset into training, validation, and test sets. Therefore Data 1 is used in fitting both these models. However, there are slightly different dataset requirements when applying statistical models for revenue forecasting, such as ARIMA and SARIMA models. These models prefer a time series data with adequate history and frequency with stationarity and no missing values. In order to fit ARIMA and SARIMA, a different dataset, Data 2, is used.

REFERENCES

[1] [Cha02] Ngai Hang Chan. Time Series Applications to Finance. John Wiley & Sons, 2002.
[2] [Gaw19] K. Gawthorpe. Random forest as a model for czech forecasting. Central European Journal of Operations Research, 27:965–977, 2019.
[3] [Kol18] A. Kolkova. The application of forecasting sales of services to increase business competitiveness. Procedia Engineering, 206:1315–1321, 2018.
[4] [LZZ18] J. Liu, P. Zhu, and J. Zhao. Short-term electricity load forecasting using random forest regression. Electric Power Systems Research, 157:35–43, 2018.
[5] [NN18] S. S. Namin and A. S. Namin. Forecasting economics and financial time series: Arima vs. lstm. Journal of Computational Science, 27:519–527, 2018.
[6] [PS19] N. Phumchusri and P. Suwatanapongched. Forecasting hotel daily room demand with transformed data using time series methods. Journal of Hospitality and Tourism Technology, 10(4):590–602, 2019.
[7] [PY19] N. Paudel and T. N. Yogi. Comparative study of machine learning algorithms for rainfall prediction - a case study in nepal. Journal of Hydrology: Regional Studies, 22:100596, 2019.
[8] [SHH17] A. Sutomo, Hasbullah, and S. Hasibuan. Comparison of arima, linear trend and single exponential smoothing for short-term forecasting of power plant spare parts inventory. Journal of Physics: Conference Series, 801(1):012072, 2017.
[9] [Tsa10] Ruey S. Tsay. Analysis of Financial Time Series. John Wiley & Sons, 3rd edition, 2010.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY