Revenue, Net Income and Gross Profitforecasting of E-Com Companies

Aniket Sarin, Sai Surya Teja Gontumukkala, Bhanu Rama Ravi Teja Gonugunta, Jeyanthi R*
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
mailto:aniketsarrin@gmail.com, saisurya029@gmail.com, ravitejagonugunta150@gmail.com, r_jeyanthi@blr.amrita.edu*

***Abstract* — In the E-commerce industry, it is critical to provide precise and dependable sales estimates. The most recent cutting-edge approaches are often univariate. Revenue forecasting is an essential component of every business plan since it allows you to determine how much and how fast you want to expand your company. Profit planning is impossible without precise profit forecasts. Profit forecasting is the estimation of future earnings after taking into account all of the elements influencing the magnitude of a company's profits, such as pricing policies, costing policies, depreciation policies, and so on. The goal of this project is to forecast revenue, net income, and gross profit data from E-Commerce companies (Amazon, Alibaba, and Walmart) using Time Series models such as Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA), as well as Deep Learning (DL) techniques such as Deep Neural Network (DNN), Bidirectional LSTM (BiLSTM), and Gated Recurrent Unit (GRU). Analyze if there is a seasonal trend or not.**

Keywords* — *ARIMA, SARIMA, DL, DNN, BiLSTM, GRU, forecasting, Time Series models, seasonal trend.

I. INTRODUCTION

Forecasting product-level demand is critical in E-commerce platforms. Accurate and dependable demand projections allow for improved inventory planning, competitive pricing, and timely promotion planning, among other benefits. While effective forecasting may result in significant savings and cost reductions, poor demand prediction has been shown to be expensive in this arena. The business climate in E-commerce is very dynamic and often turbulent, owing to holiday impacts, poor product-sales conversion rates, rival behaviour, and so on. As a consequence, demand data in this sector faces a number of difficulties, including extremely non-stationary historical data, irregular sales patterns, sparse sales data, highly intermittent sales, and so on. Furthermore, product assortments on these platforms adhere to a hierarchical structure, with some goods within a subgroup of the hierarchy being comparable or connected to one another. The time series of such connected items are interrelated and may share significant demand characteristics. For example, increased demand for one item may create a drop or rise in sales demand for another, i.e., substituting/complimentary items.

Forecasting product-level demand is critical in E-commerce platforms. Accurate and dependable demand projections allow for improved inventory management. To develop accurate and relevant projections in the E-commerce industry, it is vital to account for the concept of similarity between these items. Current E-commerce demand forecasting methods are heavily influenced by cutting-edge forecasting techniques from the exponential smoothing [1] and ARIMA [2] families. However, since these forecasting approaches are univariate, each time series is treated independently and forecasted separately. As a consequence, there are several comparable items accessible. Fortunately, a vast quantity of data is accessible in E-commerce, and this data may be used to increase prediction accuracy. Aside from historical sales data, we may gather a variety of additional log data for online commodities over time, such as page view (PV), page view from search (SPV), user view (UV), user view from search (SUV), selling price (PAY), and gross merchandise volume (GMV), among others. This information may be included into the sales prediction model utilising supervised learning approaches such as regression models, resulting in improved forecast accuracy. In general, current approaches address various series individually. It may function well in physical retail, but the quick rotation of items and the volatility of demand in online retail need the development of models that communicate information across time periods. [Yelland et al., 2010, Chapados et al., 2014, Trapero et al., 2015, Bandara et al., 2019].

Large retailers such as Walmart, Costco, Amazon, Target, and others have a distinct business strategy in which they sell both their own and rivals' items from the same location, either in-store or online. In certain circumstances, their own goods compete with those of third parties. These firms are a large data warehouse, storing data from many sources (from transactions, events, inventory to name the few). Companies like as Amazon and Walmart provide analytic assistance to third-party merchants based on the time series data they collect from each product and category, as well as useful insight from the data to optimise their commercial benefit. Every day, they have a massive volume of transactions involving many items. Analysing such enormous amounts of data and gaining useful insight is always difficult.

However, with the development of cloud computing in recent years, analysing these massive data sets in a short time period to address business challenges such as anticipating future sales and demand, product suggestion, and so on has become critical to any company's success.

Our major contributions are as follows:

- Investigated five different Time series and DL Modes namely ARIMA, SARIMA, DNN, GRU, BiLSTM for

forecasting and identify seasonal trend of E-commerce companies.

- Compared all the models and identified the best one out of them.

The paper is organized as follows: Related works were briefly and precisely discussed in Section - II. The details related to the datasets were discussed in Section – III. concepts of Time series and DL models are explained briefly in Section - IV. methodology, proposed system architecture are explained briefly in Section – V, Experimental Results were explained in Section – VI, Conclusion and Future Scope are discussed in Section – VII and followed by References.

II. LITERATURE SURVEY

Forecasting is a popular and frequently researched subject in both academia and business. Accurate forecasting substantially influences the accuracy of predicting the future, which is critical in businesses such as retail, where the uncertainty of future product sales varies dramatically. Mahalakshmi et al. [2016] provide a detailed study of time series forecasting, including prediction techniques. For time series forecasting, the general approaches described are regression methods (single and multiple variables), stochastic forecasting techniques (e.g., support vector 3 machine (SVM)), soft-computing based forecasting (e.g., artificial neural network (ANN)), and fuzzy based forecasting (e.g., fuzzy C-Means (FCM) with ANN). The techniques described above are used to anticipate electrical loads and pricing, trends and seasonality forecasting, stock selection and portfolio development, and so on. Aras et al. compared the single approach to the ensemble method. Aras and colleagues [2017]. Each forecasting approach has an advantage and works better with a given data set, for example, a linear model performs better with linear data. Facebook prophet model has piqued the interest of multiple researchers from various fields of study for predicting time series models.

Yenidogan and colleagues Yenidogan et al. [2018] compared the prophet model to ARIMA for Bitcoin forecasting and found that the prophet model outperformed the ARIMA model. The Facebook prophet model has been put to good use in the study of environmental phenomena such as ground level ambient fine particulate matter (PM2.5) concentrations. Zhao and colleagues [2018]. The authors of Zhao et al. [2018] discovered a weekly and monthly trend of PM2.5 concentrations for 220 monitoring sites in the United States between 2007 and 2015. Aguilera and co. Aguilera et al. [2019] projected daily groundwater-level (GWL) for seasonal water management and shown that the prophet model surpasses the majority of comparable methods already in use. Long Short-Term Memory (LSTM) is being used by researchers to anticipate stock market time series models. Predicting stock market values is always difficult, and Facebook Prophet does not do well in comparison to other neural network models. Mohan and colleagues [2019]. Nonetheless, Fang et al. Fang et al. [2019] proposed using LSTM and prophet to forecast the trend, followed by an inverted neural network to predict a time series, which performs better than current time series models. Another intriguing use of this prophet approach is anticipating projected meal numbers in order to plan, procure, and prepare meals for an organization's employees. Tecim and Yurtsever and Tecim [2020] demonstrated such an application of this model and argued about its accuracy and ease of usage. Wang et al. projected that the latest pandemic caused by the Covid-19 virus would peak in late October. Wang et al. [2020] used machine learning and the Facebook prophet approach in their research. Finally, relatively little work has been done on forecasting time series using tree-based algorithms.

This is primarily because tree-based algorithms predict regression by allocating the average of the leaf (i.e., the average of the training data corresponds to that leaf) to the data point being predicted James et al. [2013]. As a result, they are unable to capture the data's trend, which is prevalent in many time series data. To the best of our knowledge, the classic ARIMA model and tree-based Random Forest were employed to forecast avian influenza (H5N1) outbreaks in Egyptian chicken populations Kane et al. [2014]. For retrospective and future predictions, it has been shown that Random Forest outperforms the ARIMA model. This is due to Random Forest's ability to capture the nonlinear connection between delayed and predicted values, as well as the upward shocks of avian influenza epidemics. Motivated by this finding, we employed LightGBM, a tree-based gradient boosting approach, to estimate Walmart sales in this research.

III. DATA SET DESCRIPTION

Total Nine datasets i.e., Walmart Quarterly Revenue, Amazon Quarterly Revenue, Alibaba Quarterly Revenue, Walmart Quarterly Net Income, Amazon Quarterly Net Income, Alibaba Quarterly Net Income, Walmart Quarterly Gross Profit, Amazon Quarterly Gross Profit, Alibaba Quarterly Gross Profit are collected from Macrotrends web Site. By using web Scrapping, we converted raw data into xlsx files.

TABLE I NUMBER OF ROWS AND COLUMNS IN DATA

Parameters Considered	Walmart		Amazon		Alibaba	
	Rows	Columns	Rows	Columns	Rows	Columns
Revenue	57	2	56	2	37	2
Net Income	57	2	56	2	37	2
Gross Profit	57	2	56	2	37	2

From year 2009 to 2022 Quarterly data of Walmart and Amazon was used. For Alibaba, 2013 to 2022 Quarterly data was used.

IV. MODELS

A. *ARIMA:*

Box et al. [2015] pioneered the ARIMA modelling approach in pioneering research done in 1970. The ARIMA model takes into account three major components of historical data: autoregressive terms, moving averages, and differencing terms. And those components are often stated using models such as ARIMA (p, d, q), which states that this model employs p autoregressive terms, q moving average terms, and d differences. The ARIMA model is based on determining the structure of the data's auto-correlations function (ACF). The traditional regression model is often inadequate for understanding all of the fascinating dynamics of a time series. An ACF of the residuals of a basic linear regression, for example, exposes extra structure in the data that the

regression model cannot capture, according to Shumway et al. [2000]. Instead, the inclusion of correlation as a phenomenon that produces lags in linear relationships resulted in the creation of the autoregressive (AR) and moving average (MA) models. Box and Jenkins Box et al. [2015] introduced a non-stationary component to the AR and MA models, resulting in the commonly used ARIMA model. ARIMA is a theoretical technique that incorporates an iterative three-step model-building procedure that includes model identification, parameter estimation, and diagnostic inspection. The ARIMA model is represented by the following equation:

$$Y_t = \beta_1 + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \ldots + \Phi_p Y_{t-p}$$

where y ′ t and et denote the differenced series and the random error at time t, respectively. Furthermore, ϕ and θ are the ARIMA model parameters.

B. SARIMA:

SARIMA is also known as the seasonal difference autoregressive moving average model. SARIMA is based on the ARMA model. Its core principle is to first assess the stability of time series; then, via various phase by phase and seasonal differences, remove the regularity of non-stationary series with trend and seasonality. In a multiplicative model, the seasonal ARIMA model contains both non-seasonal and seasonal components. One abbreviation for the model is

$$SARIMA \underbrace{(p,d,q)}_{non-seasonal} \underbrace{(P,D,Q)_m}_{seasonal}$$

C. DNN:

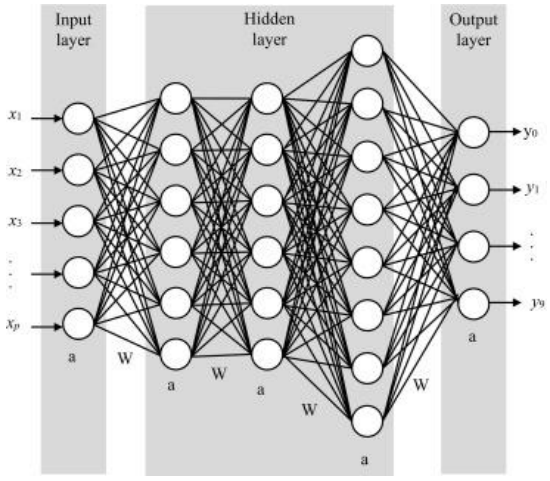

Fig.1. DNN Architecture

An ANN having numerous hidden layers between the input and output layers is known as a deep neural network (DNN). DNNs, like shallow ANNs, can simulate complicated non-linear interactions. A neural network's principal function is to accept a collection of inputs, execute more sophisticated computations on them, and output results to solve real-world problems such as classification. We limit ourselves to forward-feeding neural networks. In a deep network, we have an input, an output, and a flow of sequential data. In supervised learning and reinforcement learning challenges, neural networks are commonly employed. These networks are built on a series of interconnected levels. Deep learning may have a high number of hidden layers, most of which are non-linear. DL models outperform traditional ML networks in terms of performance. For most network optimization and loss function minimization, we employ the gradient descent approach.

D. GRU:

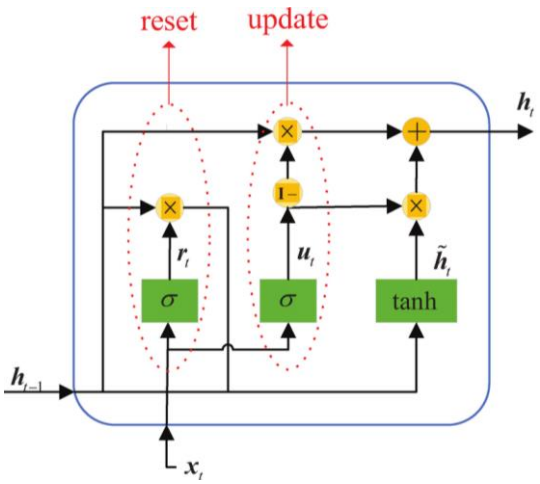

Fig.2. GRU Architecture

$$H_t = u_t \circ H_{t-1} + (1-u_t) \circ \hat{H}_t$$

It is a form of RNN that consists of two gates: a reset gate and an update gate; it functions similarly to an LSTM but with an output gate; it is favored for small data sets since it is computationally less expensive than an LSTM. A BiGRU is a data processing paradigm that consists of two GRU, one forward and one backward. BiGRU effectively enhances the amount of data accessible to the network, which improves the context that the algorithm will use. Using bidirectional with attention method will allow you to run your inputs in two directions: one from past to future and the other from future to past.

E. BiLSTM:

A BiLSTM is a sequential data processing model made up of two LSTMs, one forward and one reverse. BiLSTM effectively enhances the amount of data accessible to the network, which improves the context that the algorithm will use. When using bidirectional with attention mechanism, inputs may be routed in two directions: one from past to future and the other from future to past. A BiLSTM may aid in situations such as "Orange is a sour fruit that is round in form and orange coloured." The orange can be a colour or a fruit, and we cannot tell simply by glancing at the word "orange." It all relies on the context of the phrase. To explain a phrase in NLP, we sometimes need to relate not just to the previous word but also to the following word. In this situation, we employ forward propagation twice: once for advancing cells and once for backward cells.

$$h_t = W_{hh}\overrightarrow{h}_t + W_{hh}\overleftarrow{h}_t + b_h.$$

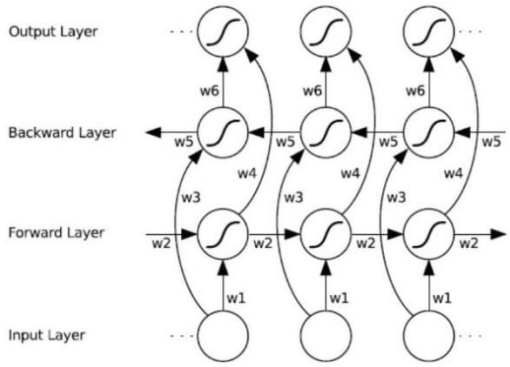

Fig.3. BiLSTM Architecture

V. MOTIVATION

Revenue forecasting is an important part of any business plan, because it can help strategize how much and how quickly you intend on growing your company. Profit planning is impossible to achieve without accurate profit forecasting. Profit forecasting means projection of future earnings after considering all the factors affecting the size of business profits, such as firm's pricing policies, costing policies, depreciation policy, and so on.

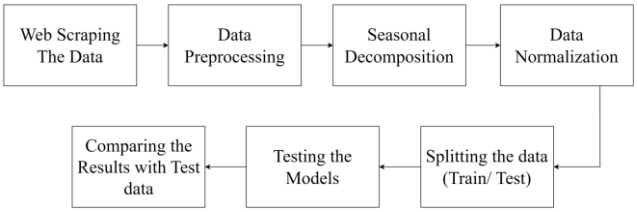

Fig.4. System Architecture

A. *Web Scraping:*

Web scraping is the practice of extracting material and data from a website using bots. Web scraping, as opposed to screen scraping, which just scrapes pixels seen onscreen, retrieves underlying HTML code and, with it, data saved in a database. After then, the scraper may reproduce the full website's content elsewhere.

B. *Data Pre-Processing:*

Time Series data contains a lot of information; however, it is not always apparent. Un-ordered timestamps, missing values (or timestamps), outliers, and data noise are all prevalent issues with time series. Handling missing values is the most complex of the challenges described. Because traditional imputation (A technique used to handle missing data by substituting missing values to maintain the majority of the information) approaches are inapplicable when dealing with time series data.

C. *Seasonal Decomposition:*

The seasonal decomposition is a time series analysis technique that represents a time series as a sum (or, in some cases, a product) of three components: the linear trend, the periodic (seasonal) component, and random residuals. The seasonal decomposition is helpful in analyzing time series that are influenced by variables that fluctuate over time in a cyclic (periodic) fashion. For example, 10-year temperature data for a certain location show a definite seasonal component, with ups and downs occurring at the same time of year in all ten years.

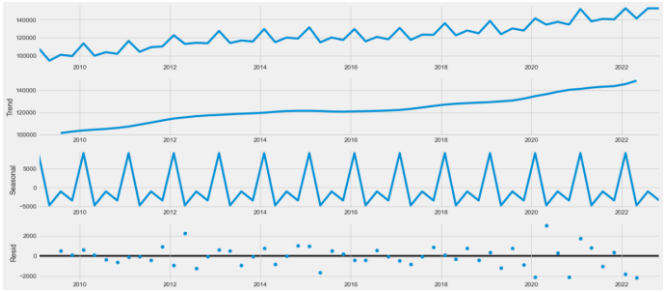

Fig.5. Seasonal Decomposition of Walmart Quarterly Revenue data

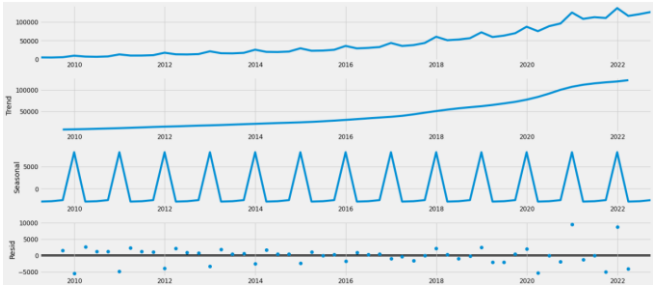

Fig.6. Seasonal Decomposition of Amazon Quarterly Revenue data

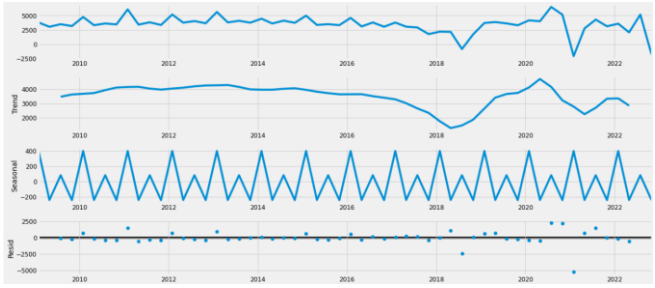

Fig.7. Seasonal Decomposition of Walmart Quarterly Net Income data

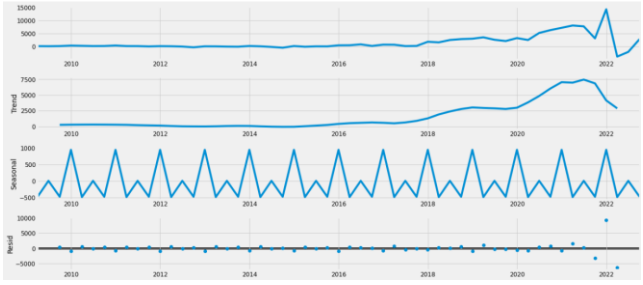

Fig.8. Seasonal Decomposition of Amazon Quarterly Net Income data

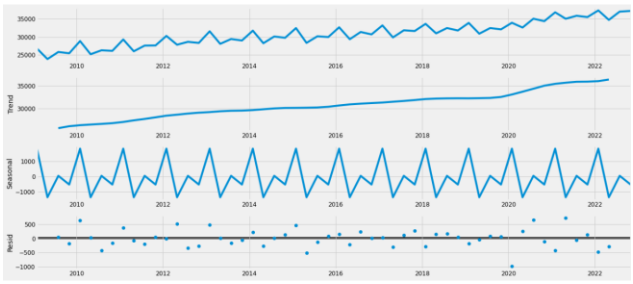

Fig.9. Seasonal Decomposition of Walmart Quarterly Gross Profit data

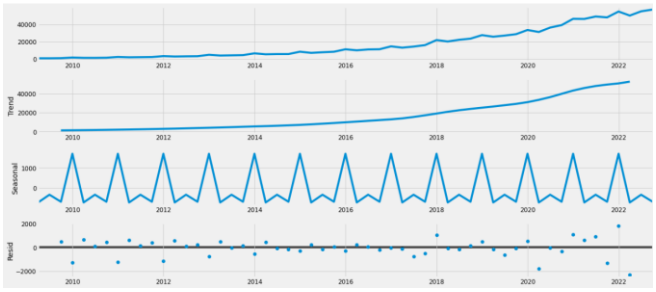

Fig.10. Seasonal Decomposition of Amazon Quarterly Gross Profit data

D. Data Normalization:

The MinMax Scaler reduces the data inside the supplied range, which is commonly 0 to 1. It scales characteristics to a specific range to change data. It adjusts the values to a given value range while retaining the original distribution's structure.

E. Splitting the Data:

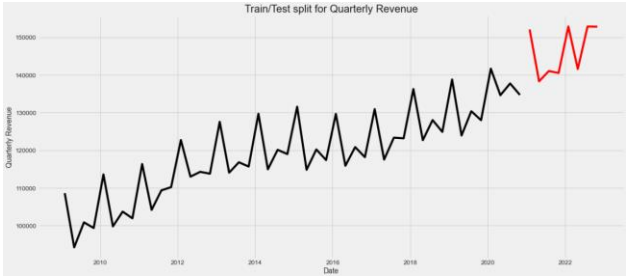

Fig.11. Splitting the Walmart Quarterly Revenue data

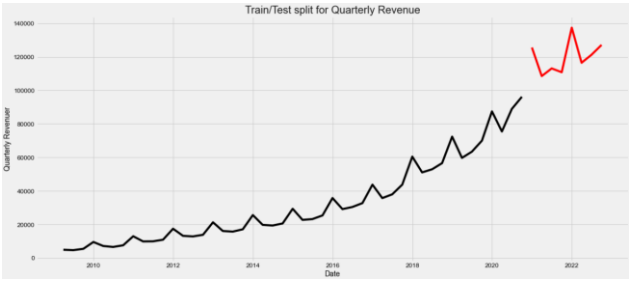

Fig.12. Splitting the Amazon Quarterly Revenue data

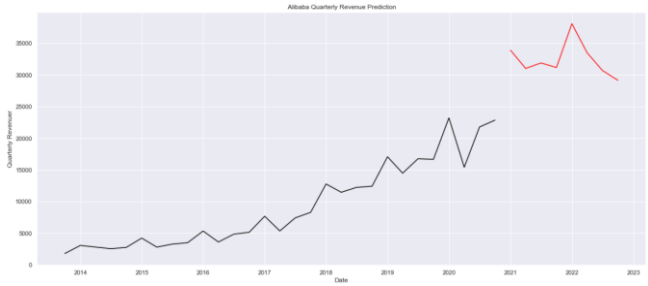

Fig.13. Splitting the Alibaba Quarterly Revenue data

VI. Results and Discussion

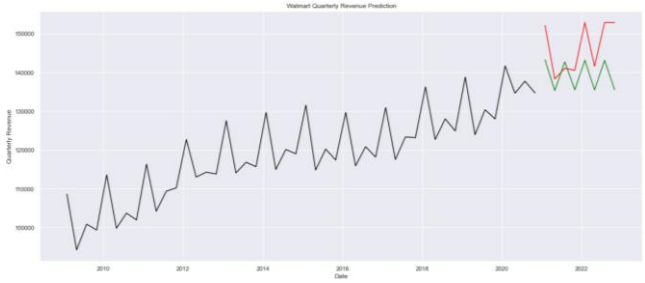

Fig.14. Walmart Quarterly Revenue Prediction for ARIMA Model

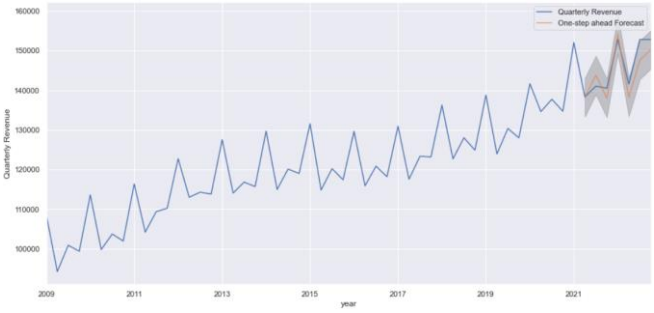

Fig.15. Walmart Quarterly Revenue Prediction for SARIMA Model

Fig.16. Walmart Quarterly Revenue Loss Prediction for GRU Model

Fig.17. Walmart Quarterly Revenue Loss Prediction for BiLSTM Model

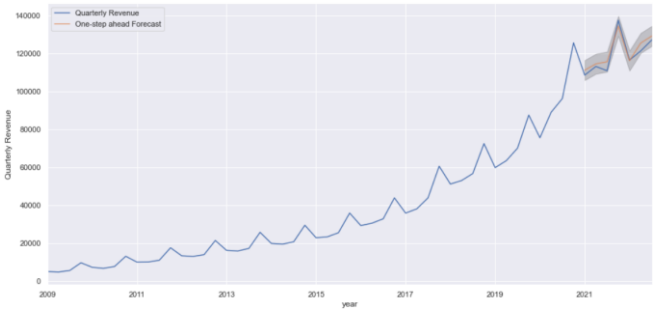

Fig.18. Amazon Quarterly Revenue Prediction for SARIMA Model

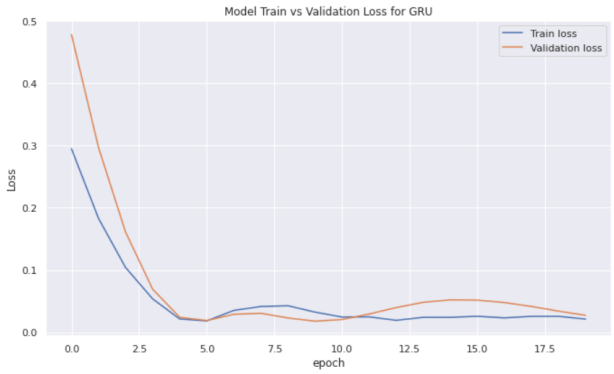

Fig.19. Alibaba Quarterly Revenue Loss Prediction for GRU Model

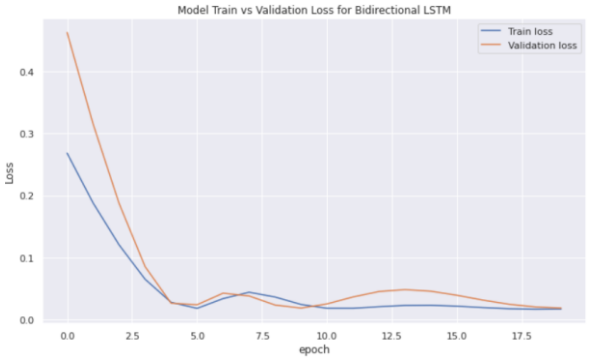

Fig.20. Alibaba Quarterly Revenue Loss Prediction for BiLSTM Model

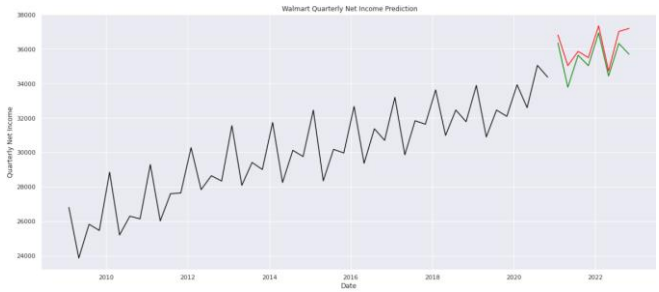

Fig.21. Walmart Quarterly Gross profit Prediction for ARIMA Model

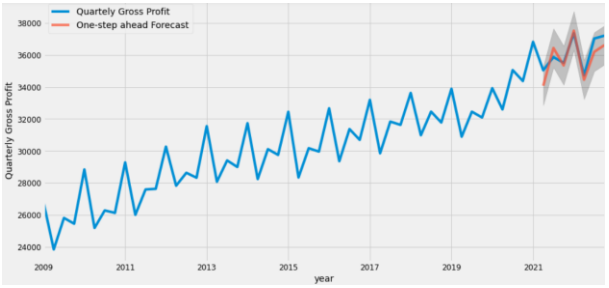

Fig.22. Walmart Quarterly Gross profit Prediction for SARIMA Model

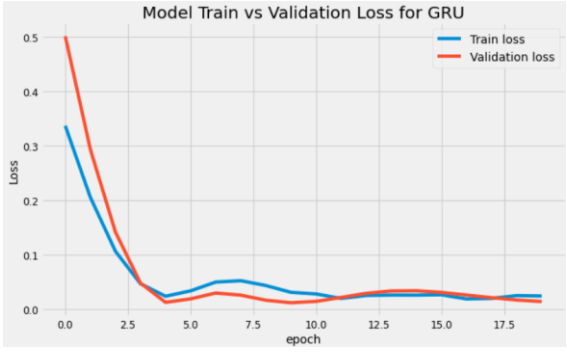

Fig.23. Walmart Quarterly Gross profit Loss Prediction for GRU Model

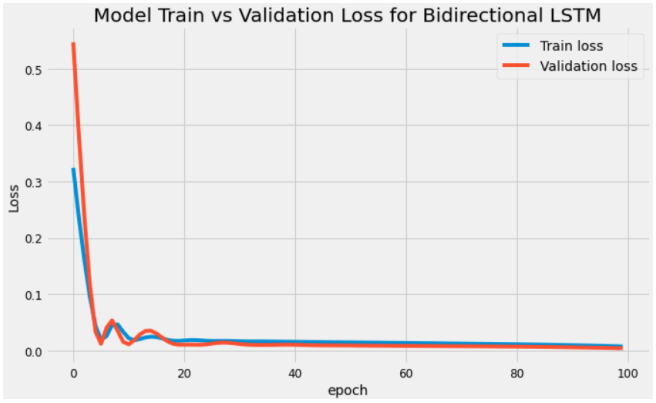

Fig.24. Walmart Quarterly Gross profit Loss Prediction for BiLSTM Model

TABLE II RMSE ERROR FOR WALMART QUARTERLY DATA

RMSE Error/ Model	Walmart Quarterly Revenue	Walmart Quarterly Net Income	Walmart Quarterly Gross Profit
ARIMA	8948.31	3373.78	795.36
SARIMA	2940.80	2561.78	587.07
DNN	5528.80	3920.88	936.02
GRU	13458.30	3120.60	2320.4
BiLSTM	9438.17	2947.60	1743.07

TABLE III RMSE ERROR FOR AMAZON QUARTERLY DATA

RMSE Error/ Model	Amazon Quarterly Revenue	Amazon Quarterly Net Income
ARIMA	14397.60	6111.10
SARIMA	2973.21	6561.76
DNN	18865.45	18333.02
GRU	25093.83	5941.35
BiLTSM	46282.99	5997.46

VII. CONCLUSION

The revenue, net income and gross profit data of E-Com Companies (Amazon, Alibaba and Walmart) were forecasted using Time Series and Deep Learning techniques. Seasonal Trend is present in the data. For seasonal data with less samples SARIMA is performing better compared to ARIMA, DNN, GRU, BiLSTM Models and Calculated the RMSE Error for all models. In future the models can be trained using larger datasets to understand deep learning models performance better.

REFERENCES

[1] Madhavi Latha Challa & Venkataramanaiah Malepati & Siva NageswaraRao Kolusu, 2020. S&P BSE Sensex and S&P BSE IT return forecastingusing ARIMA, Financial Innovation, Springer, Southwestern Universityof Finance and Economics, vol. 6(1), pages 1-19, December.

[2] Challa, M.L., Malepati, V. & Kolusu, S.N.R. Forecasting risk using auto regressive integrated moving average approach: an evidence from S&P BSE Sensex. Financ Innov 4, 24 (2018).
[3] Box, G.E.P. and Jenkins, G.M. (1970) Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.
[4] L. Menculini, A. Marini, M. Proietti, A. Garinei, A. Bozza, C. Moretti, and M. Marconi, "Comparing Prophet and Deep Learning to ARIMAin Forecasting Wholesale Food Prices," Forecasting, vol. 3, no. 3, pp. 644–662, Sep. 2021.
[5] Somarajan S., Shankar M., Sharma T., Jeyanthi R., "Modelling and Analysis of Volatility in Time Series Data". In: Wang J., Reddy G., Prasad V., Reddy V. (eds) Soft Computing and Signal Processing. Advances in Intelligent Systems and Computing, vol 898. Springer, Singapore., 2019.
[6] H. C.J., D. K.B., A. R., and J. R.," Modeling of Multivariate Systems us- ing Vector Autoregression (VAR)," Innovations in Power and Advanced Computing Technologies (i-PACT), 2019, pp. 1-6, 2019.
[7] Gamboa JC. Deep learning for time-series analysis. arXiv preprint arXiv:1701.01887. 2017 Jan 7.
[8] Lim B, Zohren S. Time-series forecasting with deep learning: a survey. Philosophical Transactions of the Royal Society A. 2021 Apr 5;379(2194):20200209.
[9] Längkvist, M., Karlsson, L. and Loutfi, A., 2014. A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognition Letters, 42, pp.11-24.
[10] Zeroual, Abdelhafid, Fouzi Harrou, Abdelkader Dairi, and Ying Sun. "Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study." Chaos, Solitons & Fractals 140 (2020): 110121.
[11] Sezer, Omer Berat, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. "Financial time series forecasting with deep learning: A systematic literature review: 2005–2019." Applied soft computing 90 (2020): 106181.
[12] Brownlee J. Deep learning for time series forecasting: predict the future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery; 2018 Aug 30.
[13] Lara-Benítez, Pedro, Manuel Carranza-García, and José C. Riquelme. "An experimental review on deep learning architectures for time series forecasting." International Journal of Neural Systems 31, no. 03 (2021): 2130001.
[14] Zhang, G. Peter. "Time series forecasting using a hybrid ARIMA and neural network model." Neurocomputing 50 (2003): 159-175.
[15] Khandelwal, Ina, Ratnadip Adhikari, and Ghanshyam Verma. "Time series forecasting using hybrid ARIMA and ANN models based on DWT decomposition." Procedia Computer Science 48 (2015): 173-179.
[16] Mehrmolaei, Soheila, and Mohammad Reza Keyvanpour. "Time series forecasting using improved ARIMA." In 2016 Artificial Intelligence and Robotics (IRANOPEN), pp. 92-97. IEEE, 2016.
[17] Domingos, S. de O., João FL de Oliveira, and Paulo SG de Mattos Neto. "An intelligent hybridization of ARIMA with machine learning models for time series forecasting." Knowledge-Based Systems 175 (2019): 72-86.
[18] Zhang GP. A combined ARIMA and neural network approach for time series forecasting. InNeural Networks in Business Forecasting 2004 (pp. 213-225). IGI Global.
[19] Li, Maobin, Shouwen Ji, and Gang Liu. "Forecasting of Chinese E-commerce sales: an empirical comparison of ARIMA, nonlinear autoregressive neural network, and a combined ARIMA-NARNN model." Mathematical Problems in Engineering 2018 (2018).
[20] Yang, Heng-Li, and Han-Chou Lin. "An integrated model combined ARIMA, EMD with SVR for stock indices forecasting." International Journal on Artificial Intelligence Tools 25, no. 02 (2016): 1650005.
[21] Zheng, Aiyun, Weimin Liu, and Fanggeng Zhao. "Double trends time series forecasting using a combined ARIMA and GMDH model." In 2010 Chinese Control and Decision Conference, pp. 1820-1824. IEEE, 2010.