2020 IEEE International Conference for Innovation in Technology (INOCON)
Bengaluru, India. Nov 6-8, 2020

Demand Forecasting for E-Commerce Platforms

Anupriya Jain
Department of ISE
BMS College of Engineering
Bengaluru, India
anupriyajain12@gmail.com

Vikram Karthikeyan
Department of ISE
BMS College of Engineering
Bengaluru, India
vikram.k.f50@gmail.com

Sahana B
Department of ISE
BMS College of Engineering
Bengaluru, India
sahana.bhaskar4@gmail.com

Shambhavi BR
Department of ISE
BMS College of Engineering
Bengaluru, India
shambhavibr.ise@bmsce.ac.in

Sindhu K
Department of ISE
BMS College of Engineering
Bengaluru, India
ksindhu.ise@bmsce.ac.in

Balaji S
Department of CSE, CIIRC
Jyothy Institute of Technology
Bengaluru, India
drsbalaji@gmail.com

***Abstract*—It gets difficult for e-commerce companies to understand market conditions. The proposed work predicts the demand for the products as per the sales in the e-commerce companies so that there is no shortage of raw materials or the number of units on the inventory side. This paper is a comparative study of the Seasonal Autoregressive Integrated Moving Average (SARIMA) model and Long Short-Term Memory network (LSTM) to predict product demand for the given dataset. Performance, scalability, execution time, accessibility and convenience are the various factors based on which the two models are compared. In SARIMA, the model with the minimum value of the Akaike Information Criterion (AIC) was selected from all the admissible models. The non-linear demand relationships available in the E-commerce product assortment hierarchy is exploited well by the LSTM**.

Keywords—Demand Forecasting, Machine Learning, Ecommerce, Time,Series, SARIMA, LSTM

I. Introduction

E-commerce retailers have found demand forecasting always challenging. Sales forecasts give the retailers a bigger picture. They get a general idea of the coming years so that they can build up their objectives to maximize their profit and success. Accurate demand forecasting can lead to an improved sale of products, storage management, business decision performance, the satisfaction of customer demands and cost reduction thereby preventing product backlog or shortage on the inventory side. Because of the rapid development of information technology in diverse fields, it has become a critical and challenging task to come up with an effective sales forecasting model as volatility of product sales has increased day by day.

In this paper, we have used two different forecasting techniques, namely, the SARIMA model and LSTM model to forecast the sales for an online store. We have implemented both the models for the superstore dataset and compared the results.

The paper is structured as follows. Literature survey for the understanding of the problem and various algorithms implemented to overcome it is presented in Section 2. Following this is the methodology section that consists of detailed information about the dataset. In the data pre-processing, the focus is mainly on the division of the dataset into training and test data along with dimensionality reduction. In the model, evaluation discussion is carried out on the statistical evaluations and also the models that are implemented are briefly described. Finally, we conclude with the result analysis for the implemented models.

II. Literature Survey

The research article by Kilimci et. al. [4] deals with Time Series and Regression methods. In this system, nine types of algorithms, namely, moving average (MA), exponential smoothening, Autoregressive Integrated Moving Average (ARIMA), Holt-Winters and three different Regression Models were employed. They have also used Support Vector Regression (SVR) which is the regression implementation of Support Vector Machine (SVM) for the prediction and classification of the continuous variables. The deep learning methods discussed here is the multilayer feedforward artificial neural network (MLFANN) in which the data and the computations flow only in one direction(forward) without any feedback. This was trained using stochastic gradient descent based on backward propagation.

The Holt-Winters machine learning algorithm is used for predicting the sales of the Walmart store in the research paper by Harsoor, Anita S., and Anushree Patil [5]. The seasonality, trend and residual randomness was observed in this algorithm and the sales prediction was done based on the training datasets

YU, Jian-hong, and Xiao-Juan LE [6] analyzed three forecasting methods on the Amazon dataset. The first was the Winters' Exponential Smoothening which took into consideration both the trend and the seasonal patterns when the smoothening process was applied. The second model was the Time series decomposition model with Box Jenkins methodology. This was used to find a linear trend of the data, along with seasonality index and a cyclical factor. The final forecasting methodology was the ARIMA where several models were compared and the model with the lowest RMSE was used for forecasting the sales. After performing sensitive analysis for the three techniques, this paper concluded that Winters' Exponential Smoothening was the least sensitive to

978-1-7281-9744-9/20/$31.00 ©2020 IEEE

changes in data while the other two were sensitive to changes in the data.

Babai, et al., [7] used the ARIMA model which exhibits a relationship between the Mean Square Error (MSE) and the inventory costs in a two-stage supply chain consisting of one retailer facing a non-stationary ARIMA (0,1,1) demand process and one manufacturer.

III. Methodology

The dataset has over 10,000 entries of product purchases over a period of 4 years, with 3 major categories. The features of this dataset are Row ID, Order Date, Order ID, Ship Date, Ship Model, Customer ID, Customer Name, Segment, Country, City, State, Postal Code, Region, Product ID, Category, Sub-Category, Product Name, Quantity, Sales, Discount and Profit.

A. Data Preprocessing

The dataset had several features some of which were either irrelevant or had insignificant effects on the demand of the products. Thus, such features were removed only retaining Order Date and Quantity sold. Then the reduced dataset was searched for any missing values and aggregated on quantity for the dates.

Since the aim was to get month-wise forecasting, the data was resampled by using average daily quantity sales for that month and using the starting of the month as the timestamp.

The processed data was divided into training and test data by date. LSTM is sensitive to the scale of data. Therefore, we have normalized the value of Quantity sold to the range of 0 to 1 using the min-max scaler. This process was skipped for the SARIMA model since it did not affect the results. Keras require input data in the form of a matrix with predicting values as one index and timestep as another dimension. The windowing method was used to convert time series forecasting to a supervised learning problem. Then the 3D vector input was prepared for LSTM which is of the form (num_samples, num_timesteps, features). This was achieved by reshaping the data.

B. Evaluation

The models are compared based on the statistical measure - Root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \qquad (1)$$

y_i is the predicted quantity; y_j is the predicted quantity with variables considered n times.

IV. Forecasting Models

This section summarizes the forecasting techniques used like SARIMA and LSTM. Once the data preprocessing is done, it is uploaded in these models which are prepared using python.

A. SARIMA

Based on dates and category, SARIMA or Seasonal ARIMA was used to train the model. It is an extension of ARIMA which also have a seasonality component. Due to this reason, SARIMA is used instead of ARIMA as data is found to exhibit seasonal pattern (repeated every 12 months). "Autoregressive" part is for the pattern of growth or decline in the data, its rate of change is accounted by "integrated" part and any noise between consecutive data points is accounted by "moving average" part. SARIMA has seven hyperparameters: - (p,d,q) - the non-seasonal part and (P,D,Q)s – the seasonal part. The general formula for SARIMA(p,q,r) x (P,Q,R,s) is given as:

$$\phi_P(B^s)\varphi(B)\nabla_s^D\nabla^d x_t = \Theta_Q(B^s)\theta(B)w_t \qquad (2)$$

Here $\Phi_P(B^s)$ is the seasonal AR component of order P, $\Theta_Q(B^s)$ the seasonal moving average operator of order Q, ∇^D_s and ∇^d are seasonal and regular differences respectively, w_t is a white noise process, φ(B) and θ(B) are the polynomials representing autoregressive and moving average components of order p and q respectively, B is backshift operator and s is the number of time steps for a single seasonal period(in our case 12 as the cycle repeats every 12 months).

The model is capable of forecasting demand of each product category for the upcoming months based on the pattern observed in the 4-year training data used.

Time series is a sequence of well-defined data measured at the uniform period over a while. It is to be noted that data collected over an irregular period does not form a time series. Meaningful statistics and characteristics of data can be found by analyzing the time series data. It can be used to find trend exhibited by data and can be quite useful for forecasting and monitoring the data points by fitting the appropriate model to it.

Since the data has many different categories and subcategories and dates can be tricky to work with, the average of daily sales for each month has been used. p, d and q values are chosen optimally using an estimator AIC. The lower the AIC value, the more optimal the solution. It is given as:

$$AIC = -2\log(L) + 2(p+q+k) \qquad (3)$$

where L is the likelihood of data and k is the intercept of the ARIMA value.

AIC estimates the quality of one model concerning other models. Finally, the RMSE value is used to predict the accuracy of the model.

B. LSTM Recurrent Neural Network

LSTM is a type of recurrent neural network architecture. Here, the short-term-memory as suggested in the name is maintained in the LSTM cell over long-time steps. It achieves this by overcoming the vanishing gradient problem. This method is advantageous as very large architectures can be trained successfully. An LSTM unit has a cell, a forget gate, an input gate, and an output gate. The cell remembers the values over an arbitrary period and the gates help in regulating the flow of information into and out of the cell.

Results

Both the above specified models were trained and compared with each other. It was found that SARIMA was better in predicting the long-term sales while LSTM gave more accurate results over short time period. Overall, the SARIMA gave us more consistent results. This section summarizes the general trends of the dataset and a comparison of their outputs.

A. *Descriptive Analysis*

Figure 1 shows the Analysis of quantity of units sold Vs time in years for the furniture category.

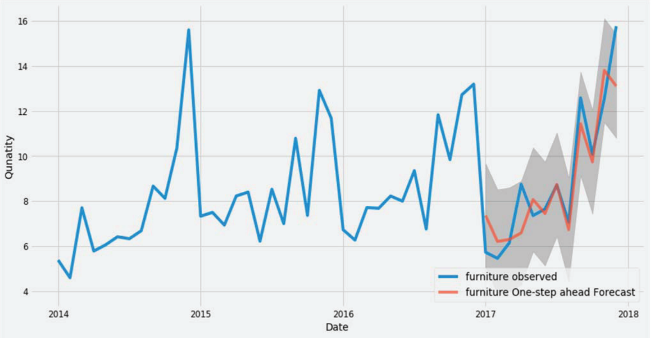

Figure 1: Analysis of quantity data vs time for the furniture category

It is observed that there is a seasonal trend, such as the quantity sold is always low at the beginning of the year and high at the end of the year. There is always an upward trend in the same year. The furniture category is taken as a whole instead of taking individual products.

B. *Forecasting Models*

The objective of all these models was to predict the number of items sold per month.

First, we built a simple moving average model with SARIMAX (1, 1, 0) x (1, 1, 0, 12) i.e. the hyperparameter value for the SARIMA model yielding the lowest AIC value 75.565. This is the most optimal option. Then we predicted the value for the next 50 months since January 2018 as shown in Figure 2.

For LSTM, the data of previous 3 months was used to predict the output of next month, i.e., given a current time (t), the prediction for the next period (t+1) is done using the current time (t) and two prior times (t-1) and (t-2).

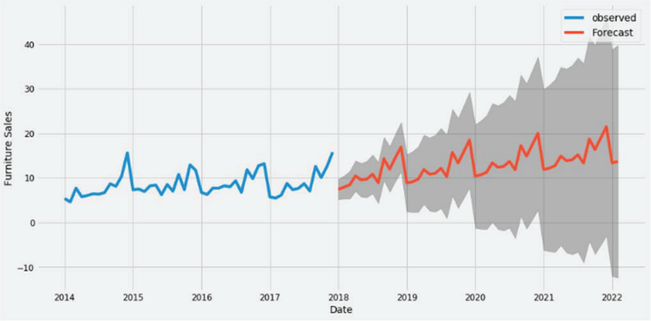

Figure 2: Quantity sold forecast for ARIMA model

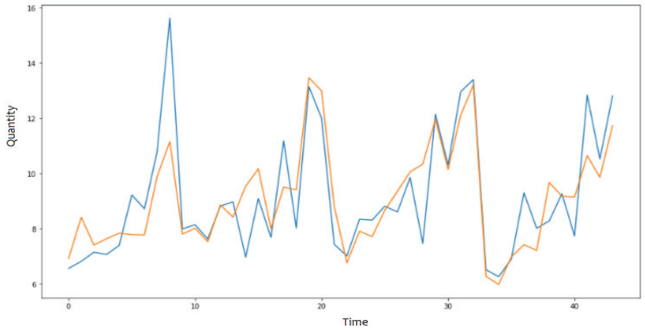

Figure 3: Validation for LSTM for training data for the furniture category.

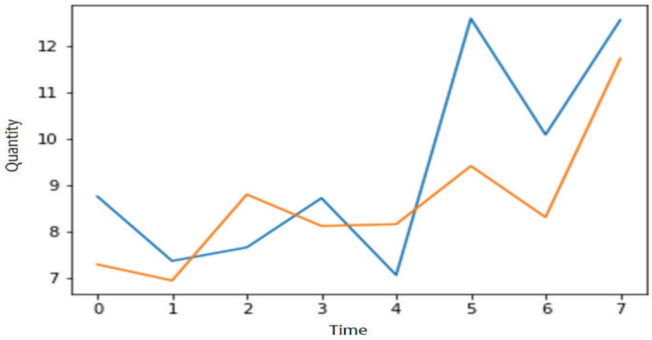

Figure 4: Prediction using LSTM for training data for the furniture category.

C. *Best Model Selection*

For retailers, it will be better if the forecast for the number of items sold is more accurate so that they can manage their inventory better. We are using RMSE values to compare the models. For the same test data of the furniture category, we found that the RMSE value for ARIMA model was 1.24 whereas for the LSTM model it was 1.55.

Conclusion

Demand forecasting is an essential component of supply chains to enhance and update stock, reduce cost, increase sales, profit and customer loyalty. Thus, models like the ones described here can be used to identify patterns, simple or complex, in historical data which may not be so apparent to us. This knowledge can then be used to predict the trends of the coming years. The advantage of forecasting is to know the number of units that can be bought by the customers to meet the production level. It helps the E-commerce platform to overcome the sales drop and also increase their profit margin. A handy digital tool driven by this model can help the E-commerce companies to avoid stock-outs and surplus units produced on the inventory side. The LSTM model and the ARIMA model are equally good for our dataset based on the RMSE values. Although the dataset contains only a few categories of products, it is possible to train the model to predict any category with appropriate training data. The performance of the model can be enhanced by using more features like shopping trends, social media response, economic studies, Location-based demographic data, items, etc. Since this model focuses mainly on the seasonal changes in demand, its best suited for common, day to day goods.

REFERENCES

[1] Jason Brownlee (July 21, 2016) Deep learning in time series.

https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras

[2] Susan Li (Jul 9, 2018) An End-to-End Project on Time Series Analysis and Forecasting with Python.

https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b

[3] Mupparaju, Kalyan, Anurag Soni, Prasad Gujela, and Matthew A. Lanham. *"A Comparative Study of Machine Learning Frameworks for Demand Forecasting."* (2018).

[4] Kilimci, Zeynep Hilal, A. Okay Akyuz, Mitat Uysal, Selim Akyokus, M. Ozan Uysal, Berna Atak Bulbul, and Mehmet Ali Ekmis. *"An improved demand forecasting model using deep learning approach and proposed decision integration strategy for the supply chain."* Complexity 2019, July 18-20, New Brunswick, New Jersey, USA

[5] Harsoor, Anita S., and Anushree Patil. "*Forecast of sales of Walmart store using big data applications."* International Journal of Research in Engineering and Technology 4, no. 6 (2015): 51-59.

[6] YU, Jian-hong, and Xiao-Juan LE. "Sales forecast for amazon sales based on different statistics methodologies." *DEStech Transactions on Economics, Business and Management* iceme-ebm (2016), June 24-26, Guangzhou, China

[7] Babai, Mohamed Ziad, Mohammad Mojiballah Ali, John E. Boylan, and Aris A. Syntetos. "Forecasting and inventory performance in a two-stage supply chain with ARIMA (0, 1, 1) demand: Theory and empirical analysis." *International Journal of Production Economics* 143, no. 2 (2013): 463-471.

[8] Dataset: https://community.tableau.com/docs/DOC-1236