Time Series Forecasing using SARIMA Model

Milan Bansal
Signal Processing and Machine Learning
Department of Electronics & Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati, India
Email: m.bansal@iitg.ac.in,

***Abstract*—Sales analysis is an important part for any working industry. It help them to keep record and predict the sales in near future. In this project report, a public data-set which includes the sales history of 4 years of a retail store is investigated to forecast the sales of furniture which is showing seasonality in its trend. To this aim, several forecasting models are applied. First, a Seasonal Auto-regressive Integrated Moving Average (SARIMA) model is applied to predict the future sales. Then a special kind of Recurrent Neural network called LSTM (Long Short-Term Memory) is applied. In addition, the data was decomposed into seasonal and non seasonal part and a SARIMA-LSTM hybrid model is applied. The performances of the models are compared using different accuracy measurement methods (e.g., Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE)).**

***Index Terms*—SARIMA, LSTM, Hybrid Model, sales forecasting, Time Series, Neural Network, seasonal .**

I. Introduction

In today's competitive world where sales are highly important for companies, accurate sales forecasting plays a key role in every successful retail business. It can be helpful in inventory management by avoiding overproduction and reducing overstock.In addition, it can be a great tool for increasing the profitability of a company by cost prediction. There are some major factors that have impacts on the future sales. These factors can be identified by analyzing the sales patterns of total sales of a retail store or a specific product. It is worth mentioning that each product has a different difficulty level of forecasting. Some products like milk, have stable consumption over the year, and therefore, their sales can be predicted easily.Some other items such as fashion goods and furniture products contain trends and seasonality in their sales pattern that cause complexity in the forecasting process.

In this study, a dataset of superstore is investigated. The products of this dataset belong to three different categories including office supplies, technology, and furniture. Among these three items, the products of furniture category are chosen to forecast their future sales. However, the historical sales of items in the furniture category of the selected dataset hold this seasonality factor. Therefore, the furniture category has been chosen for predicting the upcoming sales which can help predict the future demand. Furniture is an inseparable part of every house and office across the world.

The aim of this study is to use the past data to predict the future sales of furniture. As it has been mentioned before, seasonal times-series forecasting plays a key role in strategic decision-making and planning future activities. We also would like to check the performance of classical methods such SARIMA model and neural networks such as LSTM model for the task of seasonal sales forecasting which is widely encountered in numerous applications. In addition, we would also like to create a SARIMA-LSTM hybrid model and compare their performance.

II. System Models

The dataset contains the information of a retail store sales from 2014 to the end of 2017. The products of this dataset mainly belong to three different categories including office supplies, technology, and furniture. The dataset was first grouped on daily basis and then monthly and then divided based on category. Average monthly sales plots for each category is shown in figure.

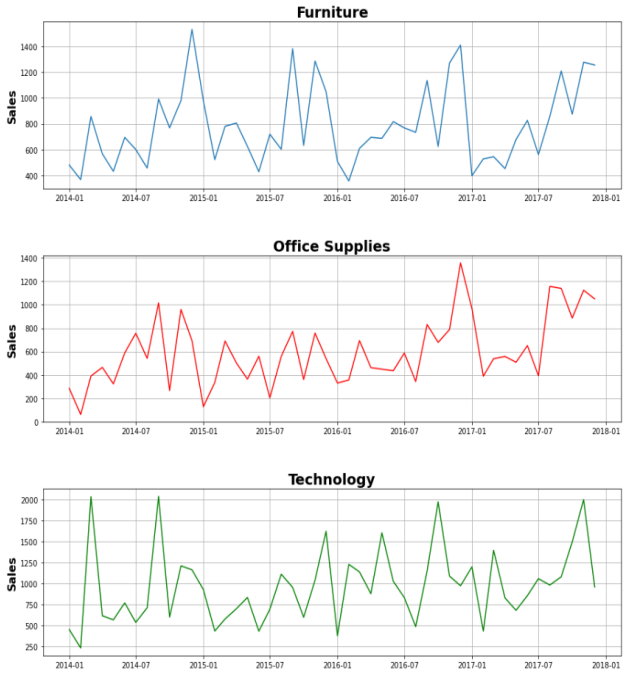

Figure 1. Sales of different Categories.

From the figure above, it can be observed that furniture sales is showing a seasonal pattern. We will only be considering the furniture sales in our study.

III. Preliminaries or Methodology or Algorithm

A. SARIMA MODEL

Seasonal ARIMA is a variation of ARIMA model. It is an extension of ARIMA method that supports seasonality in data. Let us first understand the working of the ARIMA model.

ARIMA is a statistical model used for forecasting time series data. The ARIMA equation is a regression type equation in which the independent variables are lags of the dependent variable and/or lags of the forecast errors. The equation of the ARIMA model is given as :

$$y'(t) = c + \phi_1 * y'(t-1) + .. + \phi_p * y'(t-p) + \theta_1 * \epsilon_{(t-1)} + .. + \theta_q * \epsilon_(t-q) + \epsilon_t$$

There are three terms in the equation :
AR : Auto Regression : The time series is regressed with its previous values i.e. y(t-1), y(t-2) etc. The order of the lag is denoted as p.

I : Integration : The time series uses differencing to make it stationary. The order of the difference is denoted as d.

MA : Moving Average : The time series is regressed with residuals of the past observations i.e. error $\epsilon_{(t-1)}$, error $\epsilon_{(t-2)}$ etc. The order of the error lag is denoted as q.

In the above equation, y't is the differenced series, ϕ_1 is the coefficient of the first AR term, p is the order of the AR term, θ_1 is the coefficient of the first MA term, q is the order of the MA term and ϵ_t is the error.

ARIMA does not support seasonal data. For time series that has a significant seasonal pattern, Seasonal ARIMA model is used.

In addition to the three parameters in ARIMA i.e. p, d, q, SARIMA has three more seasonal parameters (P, D, Q). The additional three parameters account for Autoregressive component (P), Differencing component (D) and Moving Average Component (Q) at the seasonal level.

It can be expressed as follows: ARIMA (p, d, q) (P, D, Q)s Here s is number of observations per season (In our case, s is 12). The seasonal components of the model are expressed in upper case and non seasonal components of the model are expressed in lower case.

The Equation for SARIMA MOdel is given as:
The general form of seasonal model SARIMA(p, d, q) (P, D, Q)s is given by:

$$\Phi_p(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \Theta_Q(B^s)\theta(B)w_t\,, \qquad (1)$$

where, w_t is the nonstationary time series, w_t is the usual Gaussian white noise process. s is the period of the time series. The ordinary autoregressive and moving average components are represented by polynomials $\theta(B)$ and $\theta(B)$ of orders p and q. The seasonal autoregressive and moving average components are $\Phi_p(B^s)$ and $\Theta_Q(B^s)$, where P and Q are their orders. ∇_d and ∇_s^D are ordinary and seasonal difference components. B is the backshift operator. The expressions are shown as follows:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - .... - \phi_p B^p$$
$$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - ... - \Phi_p B^{ps}$$
$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + .. + \theta_q B^q$$
$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + ... + \Theta_Q B^{Qs}$$
$$\nabla^d = (1-B)^d$$
$$\nabla_s^D = (1-B^s)^D$$
$$B^k x_t = x_{t-k}\ .$$

Auto-Correlation Function (ACF) : Auto-correlation of lag k is the correlation between Y(t) and Y(t-k), measured at different k lags. For lag 1, Auto correlation is measured between Y(t) and Y(t-1), similarly for lag 2, Auto correlation is measured between Y(t) and Y(t-2) values. A plot of auto-correlation for different values of k is called an auto-correlation plot or correlogram.
Partial Auto-Correlation Function (PACF) : Partial Auto Correlation of lag k is the correlation between Y(t) and Y(t-k) when the effect of all other intermediate values (Y(t-1), Y(t-2),. . . .Y(t-k+1)) is removed from both Y(t) and Y(t-k). For e.g. , partial auto correlation between y(t) and y(t+1) is the same as their autocorrelation cause there are no intermediate terms between them. Partial autocorrelation between y(t) and y(t+2) will remove the effect of y(t+1) from both y(t) and y(t+2). A plot of partial auto correlation for different values of k is called partial auto correlation plot .

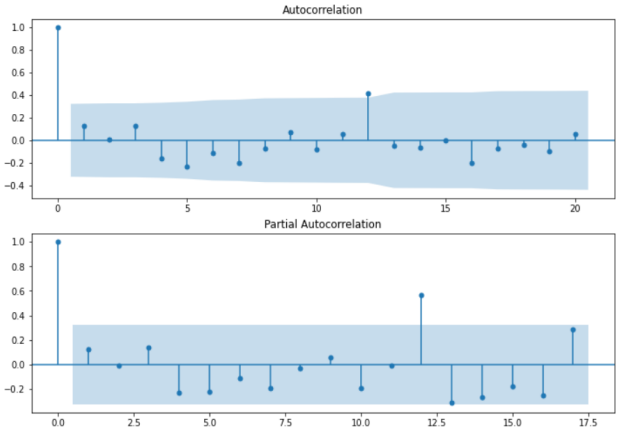

Figure 2. ACF and PACF

We then choose the best parameters from different combinations of 0 and 1 and train the model on 1st three years of the time series and predict the sales of next 11 months.The Seasonality parameter is 12.

B. LSTM MODEL

LSTM: The Long Short-Term Memory (LSTM) network is the most widely used architecture in practice to address the problem of gradient disappearance. This network structure was proposed by Sepp Hochreiter and Jrgen Schmidhuber in 1997. The idea associated with the LSTM is that each computational unit is linked not only to a hidden state h but also to a state

c of the cell that plays the role of memory. The change from c_{t-1} to c_t is done by constant gain transfer equal to 1, so that errors are propagated at previous steps without any gradient disappearance phenomenon. The status of the cell can be modified through a door that allows or blocks the update (input gate). Similarly, a door controls whether the state of the cell is communicated at the output gate of the LSTM unit. The most common version of LSTMs also uses a gate to reset the cell state (forget gate). The dynamic equations of this model are as follows:

$$F_t = \sigma(W_F x_t + U_F h_{t-1} + bF)(forgetgate)$$
$$I_t = \sigma(W_I x_t + U_I h_{t-1} + b_I)(inputgate)$$
$$O_t = \sigma(W_O x_t + U_O h_{t-1} + b_O)(outputgate)$$
$$c_t = F_t \circ c_{t-1} + I_t \circ tanh(W_C x_t + U_C h_{t-1} + b_C)$$
$$h_t = O_t \circ (tanh(c_t)$$
$$o_t = f(W_o h_t + b_o)$$

The terms b_F, b_I, b_O, bc and bo represent the different biases. Operator represents Hadamard's product (elementwise product). Initially, we take $c_0 = h_0 = 0$.

C. SARIMA-LSTM Hybrid MODEL

It is essential to control the two main components of a time Series the trend and the cyclical component. The first describes the overall movement while the second points to periodic fluctuations. SARIMA Model is effective in calculating the Seasonal component and the remaining component can be handled by LSTM. To do this, we proceed with the first step of the Box and Jenkins method, which is the $y_t = m_t + s_t + \epsilon_t$ decomposition. Three functions are obtained: the trend m_t, the seasonal component s_t and the noise or residual ϵt. s_t and ϵt are retained to reproduce seasonality and put the predicted values into confidence intervals using ϵt
. We forecast the seasonal part using SARIMA model and remaining part using LSTM model and add the results to obtain the final forecast. The flow chart is shown in figure.

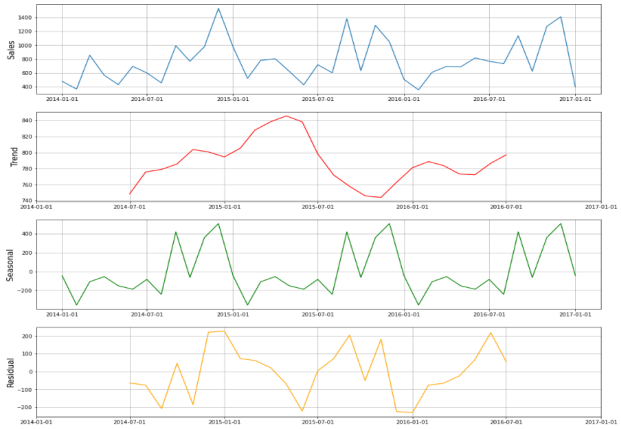

Figure 3. Components of time series

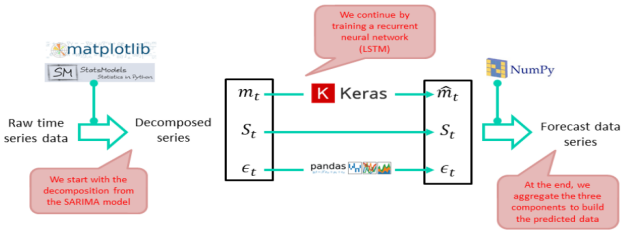

Figure 4. Hybrid-Model Flowchart

.

IV. Numerical Results and Discussions

All the Three models were trained and the forecast for next 11 months was done. The comparison is can be made on the basis of RMSE(Root Mean Square Error) and MAE(Mean Absolute Error). The Figures showing the comparison between predictions and actual data for all the three models is shown and their RMSE and MAE are calculated.

A. SARIMA MODEL

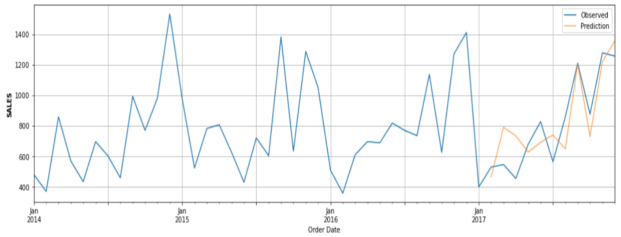

Figure 5. SARIMA Model: Actual vs Predicted

RMSE = 157.62
MAE = 133.76

B. LSTM MODEL

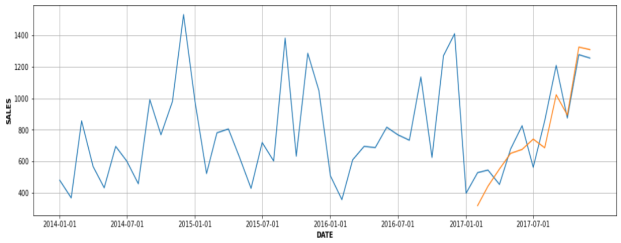

Figure 6. LSTM Model: Actual vs Predicted

RMSE = 131.13
MAE = 113.26

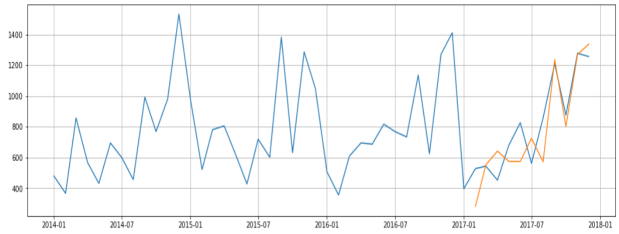

Figure 7. SARIMA Model: Actual vs Predicted

C. SARIMA-LSTM Hybrid model

RMSE = 144.13
MAE = 120.24

Comparison Table:

Model	RMSE	MAE
SARIMA	157.62	133.76
LSTM	131.13	113.23
Hybrid	144.13	120.24

D. Limitations and future scope

It can be observed from the plots that all three models performed similarly well but LSTM produced least error. It was expected that the SARIMA-LSTM hybrid model should have performed better the other two. This limitaion might have occured because here we the non seasonal part handled by LSTM part of hybrid model contained both trend as well as error. Also the data available was only of 4 years. Better predictions can be made with a larger dataset and selection of better parameters while training. These models can be applied in many sectors such as stock price prediction, electricity-load prediction, demand prediction in manufacturing industries etc.

V. Conclusions

Time Series Forecasting of furniture sales of a super store was done to predict the future sales using different models. Model appied were SARIMA Model and a neural network model LSTM. In addtion, these two models were combined make a hybrid model to work on different components of the time series : seasonal and non seasonal and combine the results to predict the future sales. It was observed, LSTM worked the best among all the three models.

References

Ruchir Kulkarni, Milind Rane, ”Pattern Recognition Product Sales Analysis Using SARIMA Model in Time Series Forecasting” in 2020 International Journal of Science and Research (IJSR), DOI : 10.21275/SR20430171526

Yuxua Han,”A forecasting method of pharmaceutical sales based on ARIMALSTM model” in 2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT), DOI:10.1109/ISCTT51595.2020.00064