SUMMER INTERNSHIP REPORT

A report submitted in partial fulfilment of the requirements for the Award of

Degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE ENGINEERING

by

Arijeet Sinha

Admn. No.: E16CSE121

Under Supervision of

Mr. Soumya Bhattacharya

ITC Infotech, Kolkata

(Duration: 14st May 2019 to 14th June 2019)

****

Department of Computer Science Engineering

School of Engineering and Applied Sciences

Bennett University

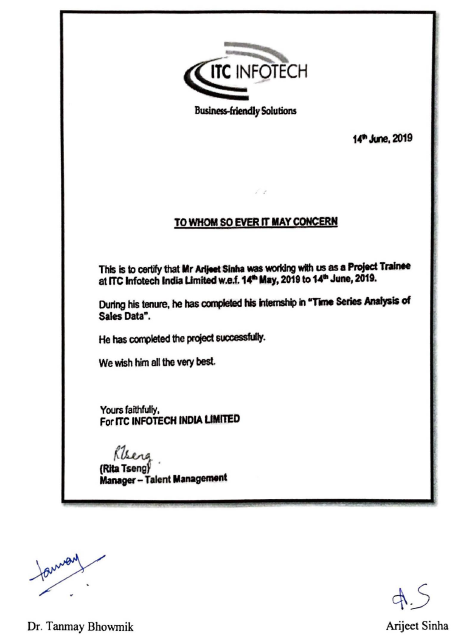
Greater Noida, Uttar Pradesh

July-2019

Internship Offer Letter



Certificate of Completion



Acknowledgement

First, I would like to thank Mr. Soumya Bhattacharya, **ITC Infotech India Limited** for giving me the opportunity to do an internship within the organization.

I would like to place my special thanks for the following

* Dr. Suneet Tuli(Dean Of SEAS, Bennett University)
* Dr. Deepak Garg(HOD Of CSE, Bennett University)
* Mr. Rakesh Singh(Head Of Career Services, Bennett University)
* Mr. Amit Krishna Biral(Project Guide, ITC)
* Dr. Tanmay Bhowmik(Internship Mentor, Bennett University)

I am extremely great full to my department staff members and friends who helped me in successful completion of this internship.

Arijeet Sinha

E16CSE121

(6thSemester, B.Tech. CSE)

Table of Contents

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Particulars** | **Page No.** |
| 1 | Abstract | 6 |
| 2 | About the Organization or Company | 6 |
| 3 | Introduction of Project | 7 |
| 4 | Related Work | 7 |
| 5 | Problem Statement | 7 |
| 6 | Proposed Solution or Approach or Technique | 8 |
| 7 | Simulation Set up and Implementation | 9 |
| 8 | Result Comparison and Analysis | 16 |
| 9 | Learning Outcome | 16 |
| 10 | Conclusion with Challenges | 17 |
| 11 | YouTube Video URL | 17 |
| 13 | Bibliography/References | 17 |

Abstract

The goal is to predict the future sales of superstore and to select a suitable model which can support seasonability of future sales and also to find the perfect period in terms of monthly, daily, quarterly or weekly. This is the study of the time series analysis of sales data from 2014 to 2018. First, analysation of data is done and then the seasonal decomposition of data and then forecasted with ARIMA (Auto Regressive Integrated Moving Average) model and Prophet model. Then selection of model is done by considering the smallest value of AIC (Akaike Information Criteria) and then RMSE (Root Mean Square Error). Here the RMSE of ARIMA model is less as compared to Prophet Model. So, ARIMA model is helpful for forecasting the seasonability data. Prophet Model is also used for finding the most suitable periods in daily and weekly.

About the Company

ITC InfoTech, which has led by Business and Technology Consulting, is a specialized global technology services provider. The company is powered by a growing portfolio of specialized solutions addressing critical business challenges, including: Industry 4.0 solutions such as PLM, IoT and Embedded; Customer Value Management; Loyalty; Trade Marketing & Distribution; Supply Chain Optimization; Data Engineering and Analytics; Healthcare Analytics; Digital Banking solutions and Digital Agriculture solutions. The company provides solutions to enterprises in Supply Chain based industries such as CPG, Retail, Manufacturing, CPG, Hi-Tech and Services such as Banking, Financial Services and Insurance, Healthcare, Airline, Hospitality through a combination of traditional and newer business models, as a long-term sustainable partner.

ITC Limited, Headquarter-Kolkata, India, is the parent organisation, which boasts of a varieties of presence in Hotels, Cigarettes, Paperboards & Specialty Papers, Packaging, Agri-Business, Packaged Foods & Confectionery, Information Technology, Branded Apparel, Personal Care, Stationery, Safety Matches.

Introduction

Time Series is defined as a series of [data points](https://en.wikipedia.org/wiki/Data_point) indexed in time order .In terms of statistics, the main aim of time series modelling is to study the past or the present observations of a time series to select or develop a suitable model for time series and to generate and plot future values for the series. There are two ways of doing it. **Time series analysis** is a method for analyzing time series data to get meaningful statistics and other characteristics of the data. **Time series forecasting** refers to the use of a model to predict future values from the previously observed values i.e. past values or present values. Time series are widely used for non-stationary data, like in economic, weather reports, stock prices, and retail sales.

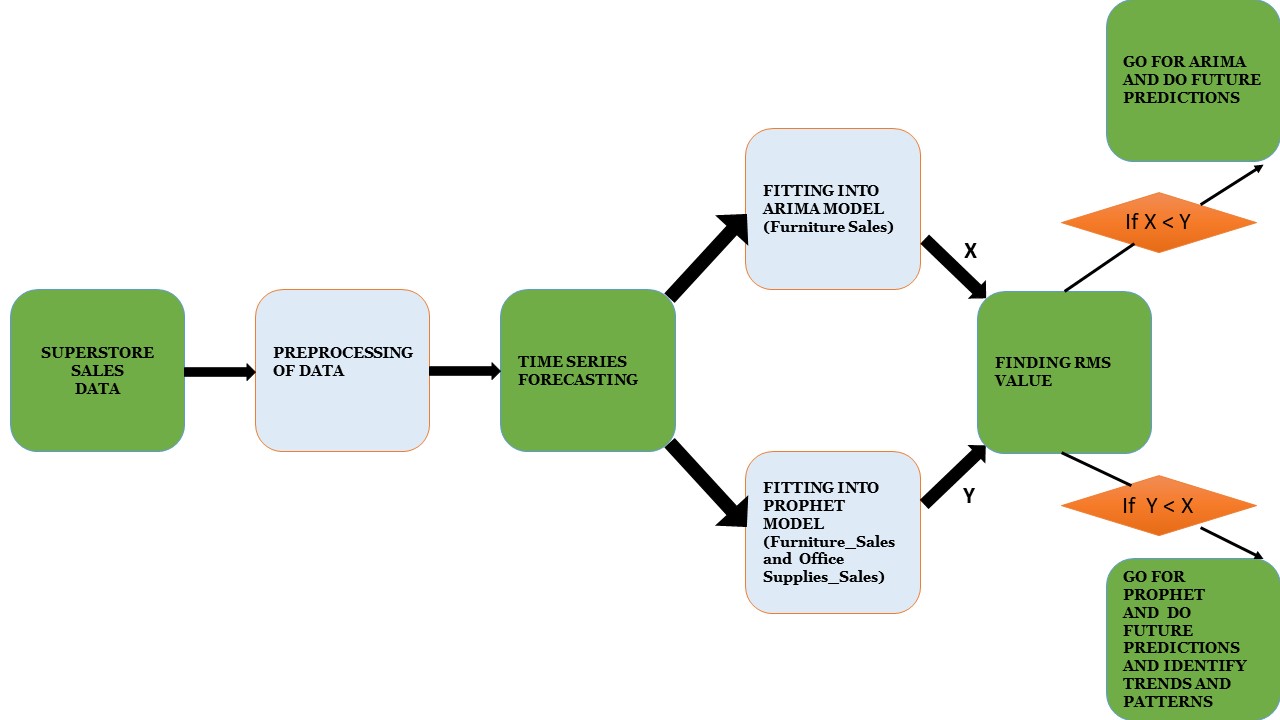
Related Work

Time Series is not a new concept and a lot of related work has been done. Here the researchers have used two different models namely, ARIMA and Prophet. ARIMA model is suitable for forecasting time series analysis of Superstore Sales Data. Prophet has been used to get monthly, weekly and yearly sales of data [3]. Ediger and Akar applied ARIMA and prophet methods to estimate the future sales from 2014 to 2018 [1]. The researchers applied ARIMA methods to predict next day sales [2]. But here I performed a comparative study of ARIMA and Prophet model for a specific time series dataset and to select suitable model for seasonal data.

Problem Statement

Predicting sales of a company requires time series data of the product and based on that data the model can predict the future sales of the products of superstore. Here, the analyzation of Time Series of Furniture sales is done and then prediction is done for the sales in the coming years for a specific product. For this Sales Predict, **ARIMA** (**Autoregressive Integrated Moving Average)** model and **Prophet Model are** applied. **ARIMA model** is applied primarily for supporting seasonal pattern of data. **Prophet** Model is applied where non-linear trends are fitted with weekly, daily and yearly seasonality, in addition to holiday effects.

Proposed Solution or Approach



**FIGURE 1**

In Figure 1, it explains the method of experiment, where the raw data is extracted from the source and some data cleaning methodology is done to make the data smooth (or consistent).

Then the Time Series Forecasting is applied for fitting into ARIMA and Prophet Model. Here X and Y denotes the RMS (Root Mean Square) Error of ARIMA model and Prophet Model, respectively.

Then selection of model is done after calculating the RMS value and then compare which model is best fitted by less RMS value.

Simulation Set up and Implementation

## **Data Ingestion & Processing**

The sample superstores sales dataset was taken from the source: **https://www.kaggle.com/pruthvi1995/superstore-sales.** It provided the sales-data from 2014 to 2018.There are several types of categories in the Superstore sales data. This dataset was used for time series analysis and forecasting, for predicting sales of furnitures.

The focus was to predict sales on of a particular order date. For that, some data pre-processing was required. This included removing columns that are not required, checking missing values, aggregate sales by-date and then indexing with time series, as current date-time data can be tricky to work with. Therefore, instead of month, calculation of daily sales in done for that month, and start of each month is used as the timestamp.

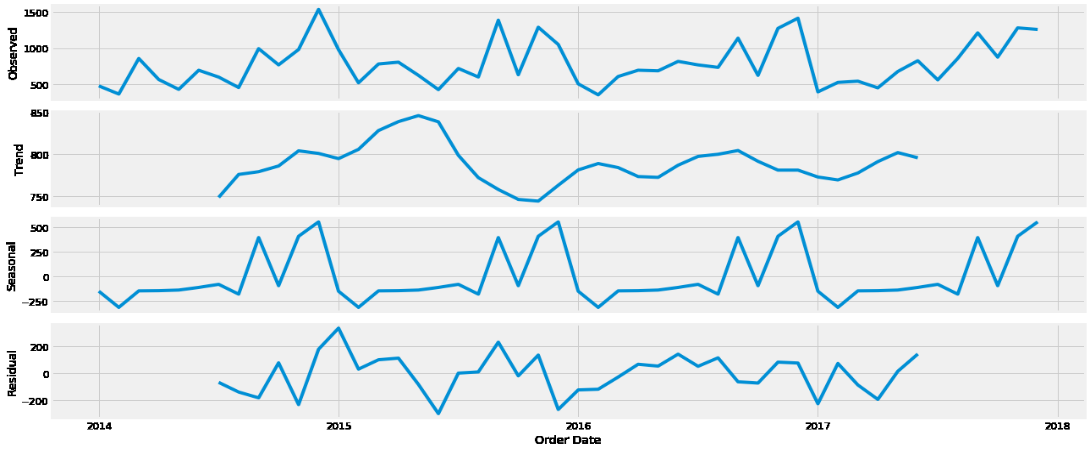
## **Time Series Analysis**

Here, visualisation of furniture sales time-series data is done and some distinguishable patterns appear while plotting the data. A look at Figure 2, shows there is always a strong upward trend within any single year with a couple of low months in the mid of the year. The sales are always low at the beginning of the year and high towards the end of the year at time series seasonal pattern.



**FIGURE 2**

Time-series decomposition is used, which allows to decompose time series into three distinct components of seasonal pattern: trend, seasonal, and residual. In Figure 3, the plot clearly shows that the sale of furniture is unstable, along with its obvious seasonality.



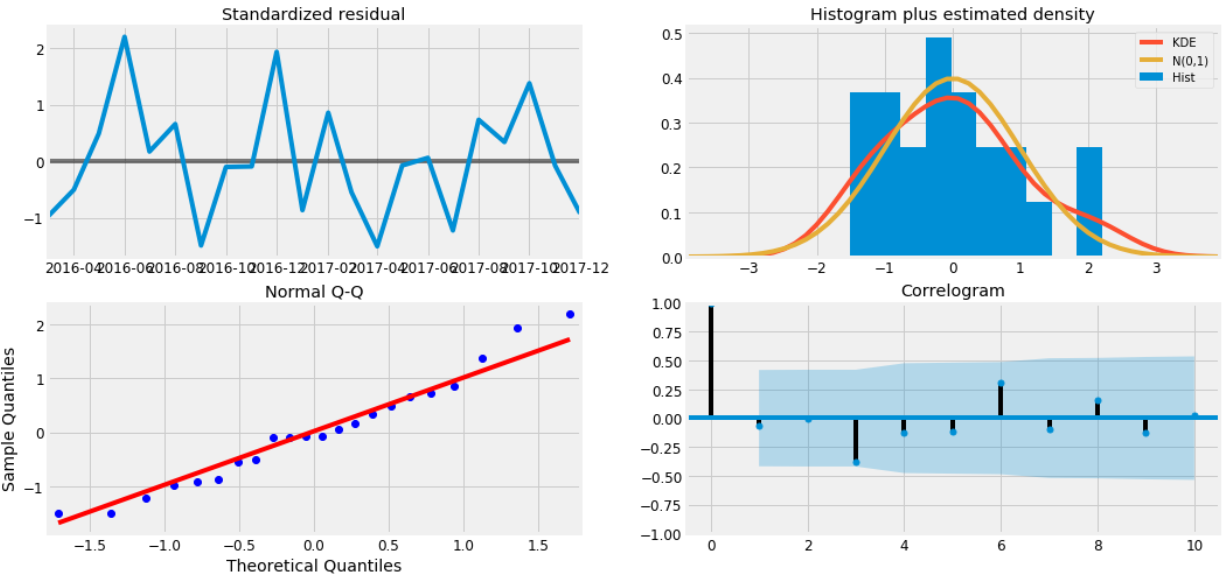
**FIGURE 3**

## **Time Series forecasting with ARIMA model**

In ARIMA model, the function SARIMAX (p, d, q) m is implemented. These four parameters account for seasonal (p), difference (d), MA (q), and the number of time steps for a single seasonal period (m) in data. By default, m =12 as it is for monthly data for a yearly seasonal cycle, so grid search is applied in order to find the optimal parameter that gives the best performance of the model

**AIC = -2log (L) + 2(p+q+d),**

Where *p* is the order of the autoregressive part and *q* is the order of the moving average part and L is the likelihood of the data and *k* as the intercept of the ARIMA model. Till now, the output of SARIMAX (1, 1, 0, 12) yielded the lowest AIC value of **297.7875439553055**. Therefore, this value is considered to be an optimal option. Model Diagnostics is also done to check for any unusual behavior. In figure 4, the behaviors are not perfect, however, model diagnostics suggests that the model residuals are near normally distributed.

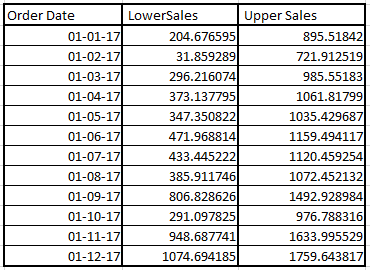


**FIGURE 4**

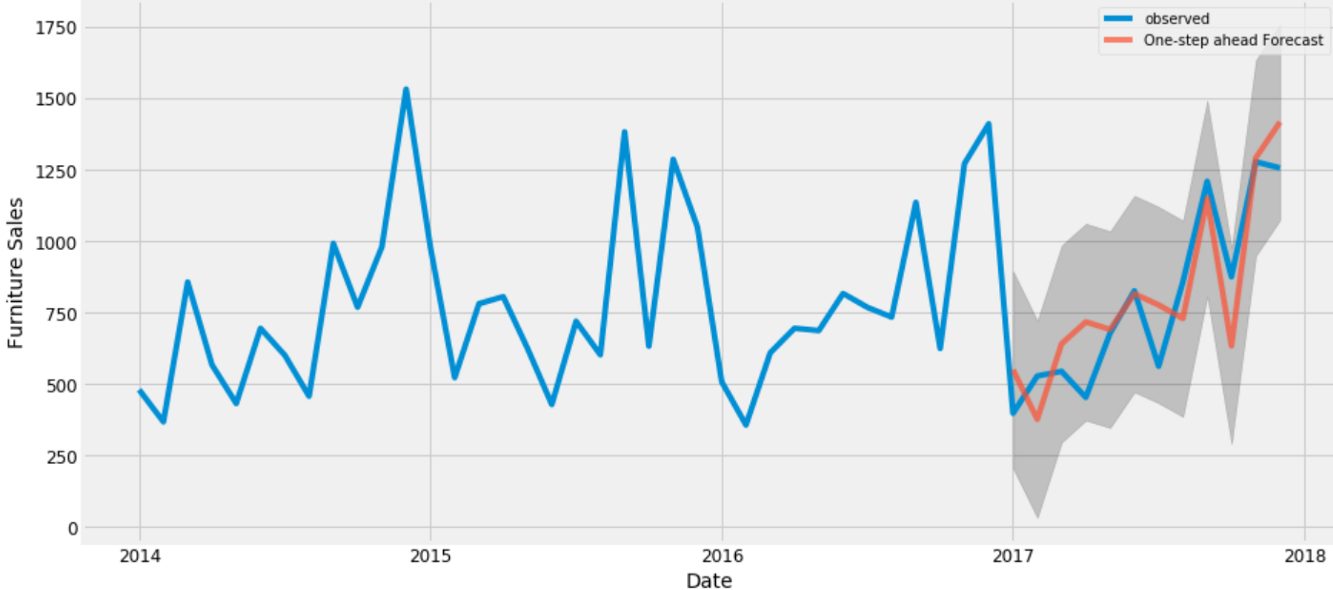
1. **Validating Forecast**

Predicted sales is compared to real sales of the time series, and set forecasts to start on 2017–01–01 to the end of the data, in order to get forecasts’ accuracy. The Table 1 shows the confidant interval values from 2017-2018.

In Figure 5, the line plot is showing the observed values compared to the rolling forecast predictions from 2017 to 2018. Overall, there is a better alignment of futures values and true values, indicating an upward trend started from the beginning of the year and at the end of the year captures seasonability.



**TABLE 1**

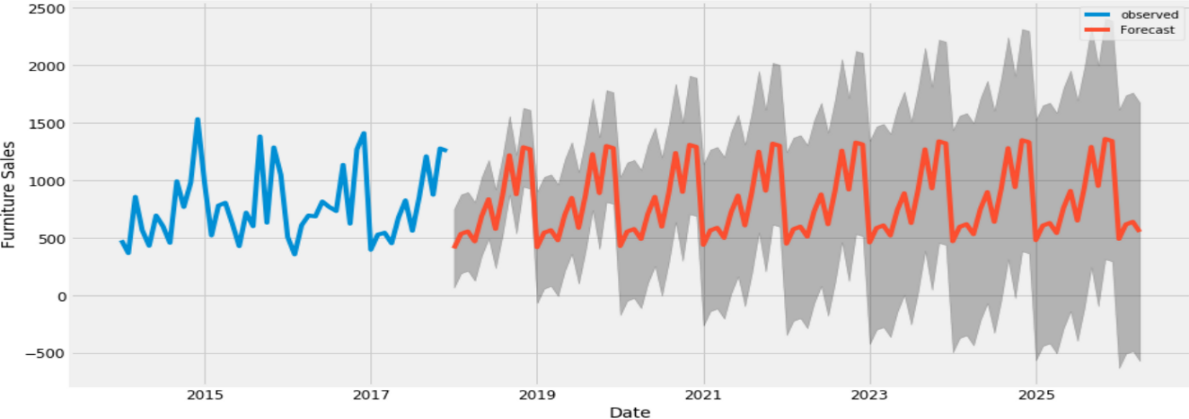


**FIGURE 5**

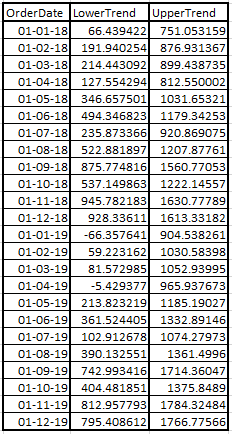
## **Predicting Future Sales**

Here, future sales is compared to the predicted sales of the time series. Then the future forecasts sets to start on 2018–01–01 to the end of the data. The Table 2 gives the confidant interval values from 2019 till so on.

In Figure 6, my model clearly captured furniture sale-seasonality. Further forecasting is done into the future values and it became less conf**i**dant in values. This is due to these confidant intervals generated by the model, which grows larger as move further into the future.



**FIGURE 6**



**TABLE 2**

## **Time Series of Supplies**

The time series analysis for furniture makes me eager to know about other categories, and how do they compare with each other, over time. Therefore, time series of furniture and office suppliers is compared. According to data, the Office Supplies clocked 3,905 sales more than Furniture over the years (Furniture - 2121, Office Supplies – 6026).

So particular order date is focused and same procedures is applied for office supplies as that of Furniture. Some data pre-processing is done and that included removing columns that’s not needed, check for the values which is missing, sales by-date aggregating and so on. Also time series indexing is done. Instead of the start of each month as the timestamp, average daily sales value is used.

Then comparision of two categories’ is done in sales in the same time period. Therefore, two data frames are combined into one and two categories’ time series are used for plotting into one plot.

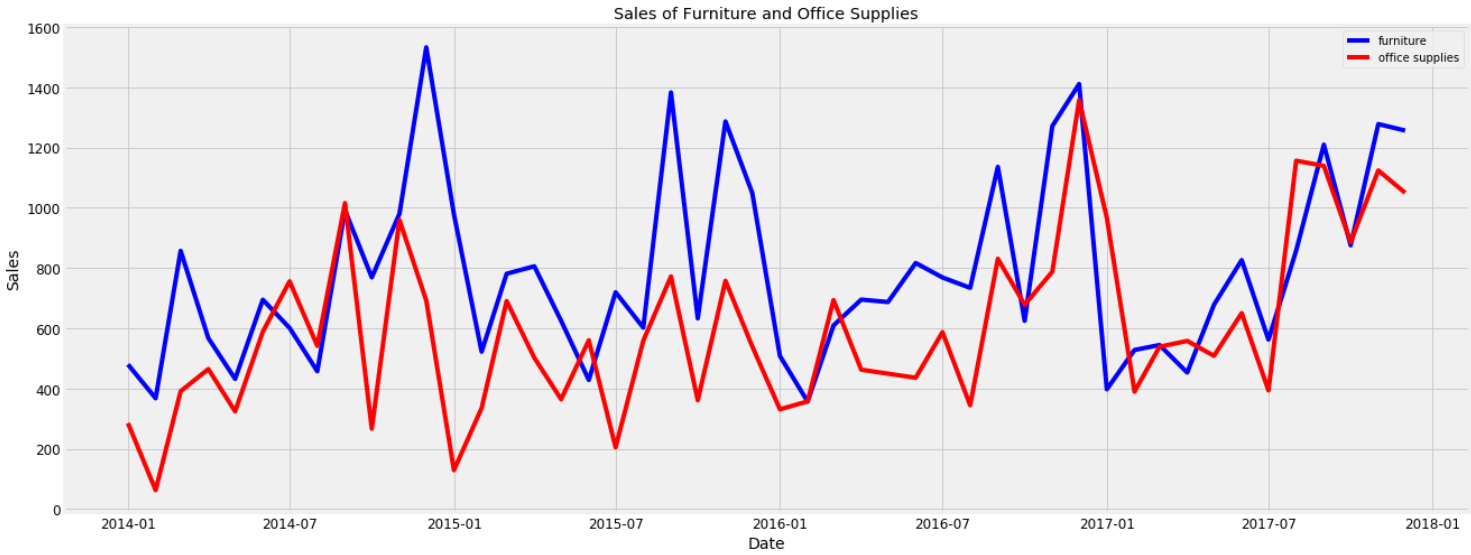
## **Time Series Modelling with Prophet**

Here the sales of furniture and office supplies have a similar seasonal pattern, as shown in Fig 7. It seems summer time is quiet for office supplies. The average daily sales for furniture are higher than those of office supplies in most of the months. Therefore, the value of furniture should be much higher than the value of office supplies. Occasionally, office supplies surpassed furniture on average daily sales.

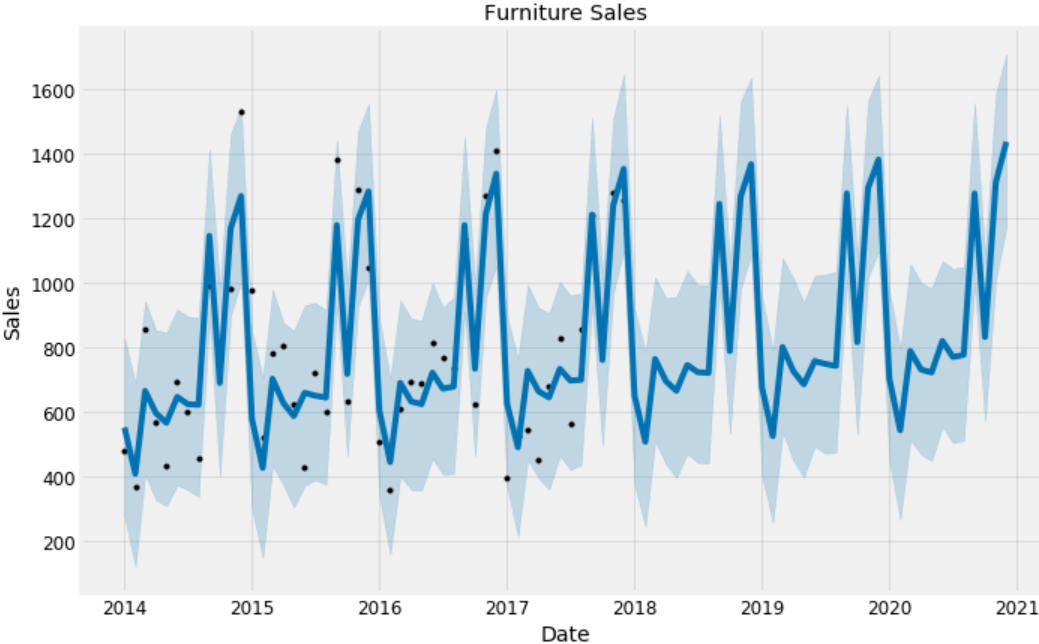
Therefore, Prophet is used to get a model for execution. Figure 8 shows the future predictions of furniture and Figure 9 shows the future predictions of office supplies.

**y(t) = g(t) + s(t) + h(t),**

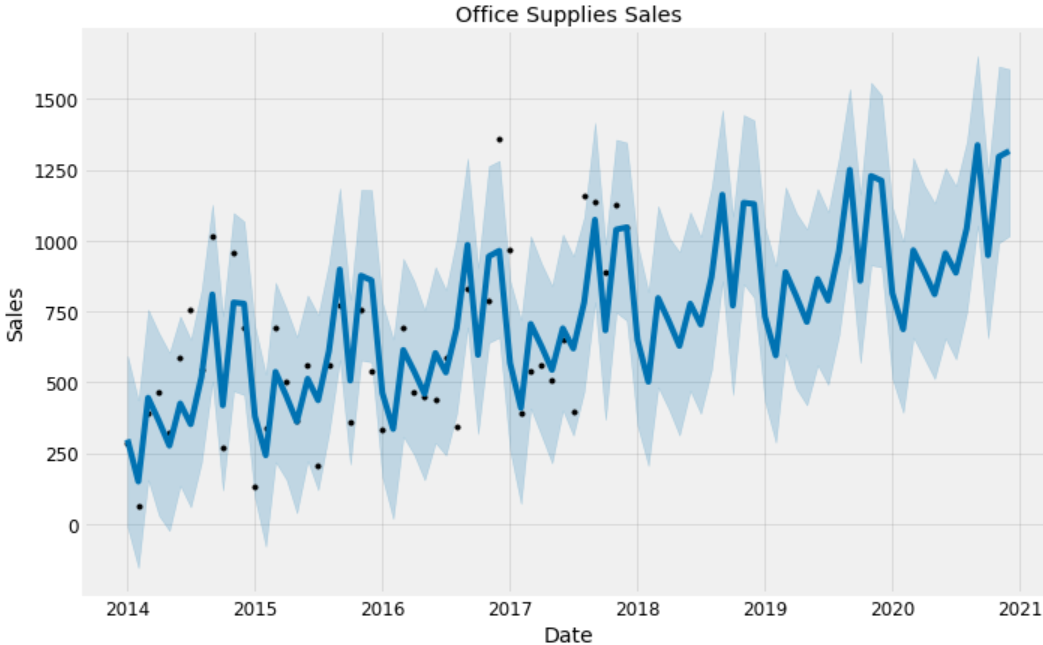
Where **g (t)**: piecewise linear growth curve for modelling non-periodic changes in time series, **s (t)**: periodic changes (yearly, monthly, weekly seasonality)**, h (t)**: effects of holidays that user has provided with irregular schedules



**FIGURE 7**

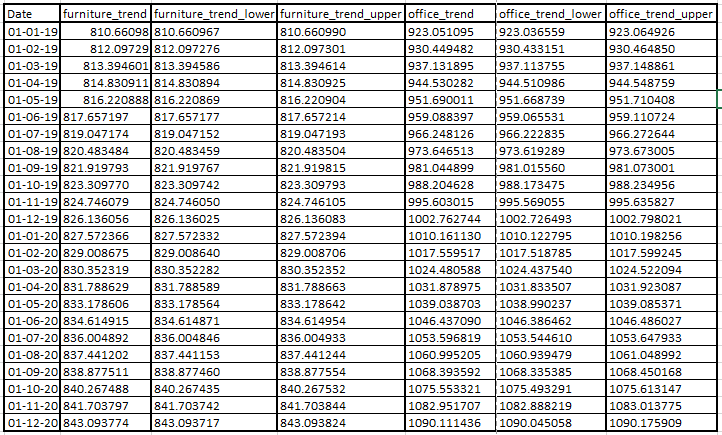


**FIGURE 8**



**FIGURE 9**

Forecasts for 2019 to 2020 is already done for these two categories. Then two categories is joined to compare their future forecasts as shown here in Table 3.

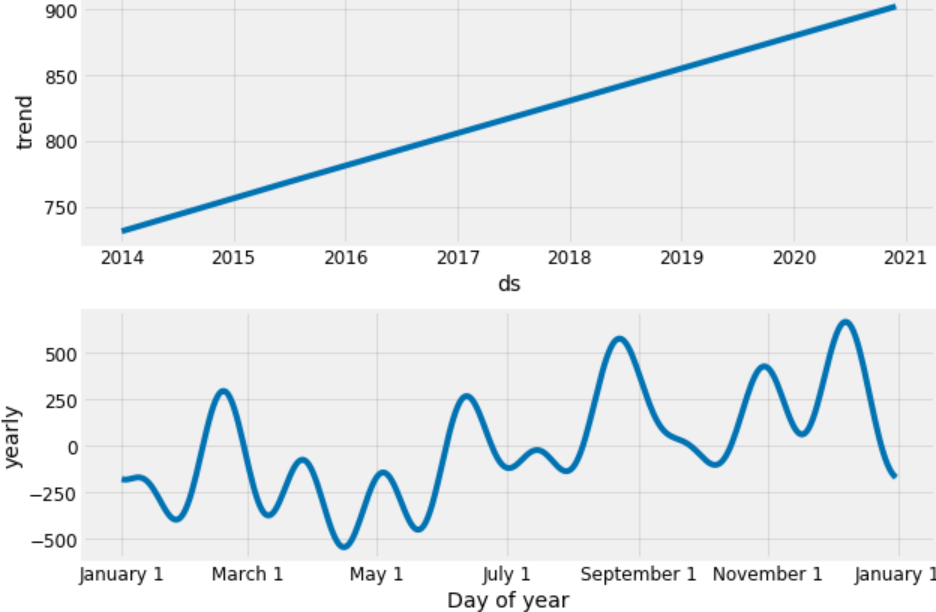


**TABLE 3**

## **Trend and Patterns Visualization**

Now, Prophet model can inspect different trends of these two categories in the data. Although office supplies’ growth seems slightly stronger, the sales for both furniture and office supplies have been linearly increasing with time and in future will keep growing, as shown in Fig 10.

It is observed that April is the worst month for furnitures and December is the best month for furnitures. February is the worst month for office supplies and October is the best month for office supplies.



**FIGURE 10**

Result Comparison and Analysis

MSE is calculated by taking the average of the squares of the errors. MSE is always positive and lesser the MSE value, more chance of finding the line of best fit

Root Mean Square error is measured as the differences between estimated values and predicted values of the model. In general, a lower RMSE value of the model is better than the higher RMSE value of the model

**MSE = (1/n)2 ,**

where Yi is the actual value and Yi’ is the predicted value.

**RMSE = √MSE**

ARIMA model has an RMSE value of 151.64 and was able to forecast the average daily furniture sales in the test set within RMSE value of the real sales. Prophet model has an RMSE value of 786.77 and was able to forecast the average daily furniture sales and office sales in the test set within RMSE value of the real sales.

Therefore, ARIMA model is suitable for a perfect fit to the date as its RMSE value is lower than RMSE of Prophet Model. Moreover, Prophet Model won’t be able to show monthly and weekly pattern of data.

There are many other time-series analysis techniques with which one can predict, such as forecast with uncertainty bounds, anomaly detection and change point, forecast time-series with other external data source.

Learning Outcome

Through my 5 Weeks summers at ITC Infotech, Kolkata, I have been able to demonstrate how the Future Sale of Office Supplies and Furniture may be affected due to the vagaries of seasons. The basis of extrapolation was using the current or Present Data of Sales

I have learned the following techniques and features of a time series data set

- Time Series Forecasting and Analysis

- Moving Average

- Seasonal Availability of Products

- Stationary or Static Movements of Certain Products off the shelves during Slack Season

Conclusion with Challenges

* My goal is to predict the future sales of superstore and to select a suitable model which can support seasonability of future sales and also to find the perfect period in terms of monthly, daily, quarterly or weekly.
* Here the RMSE of ARIMA model is less as compared to Prophet Model. So, ARIMA model is helpful for forecasting the seasonability data.
* Prophet model is used for just for plotting periods in daily and yearly.
* Analysation is only done for dataset from 2014 to 2018. If we get more past data is analysed, then we can get and forecast deeply.
* RMS error can be decreased by deep learning models other than LSTM.

YouTube Video URL

<https://www.youtube.com/watch?v=yAyTb8VNld4&feature=youtu.be>

Bibliography/References

[1] Volkan Ş. Ediger, Sertaç Akar, “ARIMA forecasting of furniture sales,” Energy Policy, vol.35, 2007, pp. 1701-1708.

[2] Javier Contreras, Rosario Espinola, Francisco J. Nogales, and Antonio J. Conejo, “ARIMA models to predict next-daysales” IEEE Transactions on 2003, vol.18, no. 3, pp. 1014- 1020.

[3] Antonio J. Conejo, Miguel A. Plazas, Rosa Espinola, and Ana B. Molina, “Day-ahead sales forecasting using the ARIMA models,