INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
ALLAHABAD

6^{TH} SEMESTER MINI PROJECT

Sales Prediction using ARIMA Model

Submitted By:
Komal Goenka (IIT2016144)
Rahul Udaiwal (IIT2016142)
Kaniskha Shekhar (IIT2016003)
Mandeep Chakma (IIT2016012)
Himanshu Singh (RIT2015049)

Submitted To:
Dr. Ranjana Vyas
Assistant Professor

CANDIDATE'S DECLARATION

We hereby declare that the work presented in this project report entitled **"Sales Prediction System using ARIMA Model",** submitted towards fulfillment of 6th semester (2019) of B.Tech (IT) at Indian Institute of Information Technology, Allahabad, is an authenticated record of our original work under the guidance of **Dr. Ranjana Vyas**. Due acknowledgements have been made in the text to all other material used.

Date : 03/05/2019
Place : Allahabad

CERTIFICATE FROM SUPERVISOR

I hereby recommend that the mini project report prepared under my supervision, titled **”Sales Prediction System using ARIMA Model”,** be accepted and the above statement made by the candidate is correct to the best of my knowledge.

Date : 03/05/2019	Supervisor:
Place : Allahabad	**Dr. Ranjana Vyas**

Abstract

Our work will contribute to modelling and forecasting the demand in a Superstore. We aim to demonstrate how the historical demand data could be used to forecast future demand and how they affect supply chain. Time Series Analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data.Time Series works for non-stationary data like economic, weather, stock price, and retail sales like in our dataset.Here we will be using Autoregressive Integrated Moving Average (ARIMA) model. The model would correspond to the standard ARIMA(p,d,q) which will be validated by another historical demand information under the same conditions. The proposed solution uses an univariate time series model for future prediction of sales. The results obtained will prove that the model could be utilized to model and forecast the future demand. Released by Facebook in 2017,we will be using forecasting tool PROPHET, which is designed for analysing time-series that displays patterns on different time scales such as yearly, weekly and daily.It also has advanced capabilities for modelling the effects of time-series and implementing custom change points.

In our model, the first step that is required is to convert our time series data into regular time series data. And after that, by taking different values of non-seasonal (p,d,q) and seasonal (P,D,Q) parameters, we train our model. We calculate Akaike information criterion (AIC) or Bayesian information criterion (BIC) value for each model and check which model best fits to our training dataset and then calculate the error. Basically this model is decomposing our original plot into patterns and randomness and on the basis of that we are predicting out future sales.

Contents

1 Introduction

In today's world where competitive margins are becoming increasingly narrower and actions must be decisive yet informed, the ability to accurately make forecasts is of premier importance. To respond quickly to shifting demand, organizations are moving toward a more effective demand-driven supply chain. Sales Prediction is crucial to inventory management as well as increasing population demand. In fact, inaccurate estimation of sales can cause significant costs to pay, which proves that the process is not improved. Consequently, many systems incur large investments in inventories to avoid "stock outs." A further complicating issue is that some demands can be intermittent demands, which means that there is a time when we have no demand and other time when we have successive demands. If a company can match the demand of a product with just the right amount of supply, then there will be no lost sales due to a lack of inventory as well as no costs from overstocking. Sales forecasting uses patterns gleaned from historical data to predict future sales, allowing for informed courses-of-action such as allocating or diverting existing inventory, or increasing or decreasing future production. Determination of parameters used in the model requires good intuition, which comes from experience, ability to identify randomness, seasonality, patterns, by just having a glance at the actual plot of the data and then running the diagnostics to see the statistics, which determine the quality of the model.

In this project report we have described the time series model and its implementation for for the prediction of future sales. This is implemented using Autoregressive Integrated Moving Average (ARIMA) model.

2 Motivation

In today's organizations, business sector need more accurate and practical reading in future to be successful. The forecasts are becoming very crucial since they are the sign of survival and the language of business in the world. Forecasting is the operation of making assumption about the future values of studied variables.Sales Prediction is crucial to inventory management as well as increasing population demand. In fact, inaccurate estimation of sales can cause significant costs to pay, which proves that the process is not improved.

To overcome such troubles, we here aim to implement a time series sales forecasting model using Box Jenkins methodology. Our basic motivation is that if we have a rough prediction of our future sales then we can recognize our challenges and be prepared for them by taking necessary actions before hand. Also we can figure out some useful business strategies to improve our sales.

Analysis may include, suppose a particular product's sale is significantly large in January(As predicted by this project) then we can maintain it's stock quantity and offer discount when the sales is low.

3 Problem Definition

The problem consists of developing a model that is able to forecast future sales quantities depending on the given sales history. Predicting the sales of a superstore needs sales history data of that superstore and based on that data, the model can predict the future sales of that superstore or product by generating trends, seasonality, residuals etc. For this project of sales prediction, we will apply ARIMA model of Time Series Analysis and evaluate the results based on training, testing and validation of the data set.

4 Literature Review

In today's organizations, which are subject to abrupt and enormous changes that affect even the most established of structures and where all requirements of business sector need accurate and practical reading into future, the forecasts are becoming very crucial since they are the sign of survival and the language of business in the world [1].

Any forecast can be termed as an indicator of what is likely to happen in a specified future time frame in a particular field. Therefore, the sales forecast indicates as to how much of a particular product is likely to be sold in a specified future period in a specified market at specified price.In sales prediction, forecasting demands is among the most crucial issues in inventory management, to increase the performance of the sales and set the sales target, to plan the budget and predict the future expenses, to identify the important segments of customers, to estimate the amount of sales required at particular periods of time.

Forecasting method involves different model: one of the important model is time series. A time series is nothing but observations according to the chronological order of time.Time series forecasting models use mathematical techniques that are based on historical data to forecast demand. It is found on the hypothesis that the future is an expansion of the past; that's why we can definitely use historical data to forecast future demand [1].

One of the most important and widely used time series models is the autoregressive integrated moving average (ARIMA) model [3]. The popularity of the ARIMA model is due to its statistical properties as well as the well-known Box–Jenkins methodology [6] in the model building process.These models are linear since the future values are cramped to be linear functions of past data [1].Box and Jenkins [6] developed a practical approach to building ARIMA models,which has the fundamental impact on the time series analysis and forecasting applications. The Box–Jenkins methodology includes three iterative steps of model identification, parameter estimation and diagnostic checking [3].This three-step process is repeated several times until a satisfactory model is finally selected. Then this model can be used for forecasting future values of the time series.

5 Proposed Methodology

5.1 Understanding Time Series

Forecasting method involves different models: one of the important models is Time Series Analysis. The first step in the Sales Prediction System is to understand what is time series. A time series is nothing but observations in chronological order of time. It is a sequence of observations taken sequentially in time. Time series forecasting models use mathematical techniques that are based on historical data to forecast demand. It is found on the hypothesis that the future is an expansion of the past; that's why we can definitely use historical data to forecast future demand. Time Series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values.

5.1.1 Components of a Time Series

- **Trend :** Trend is a general direction in which something is developing or changing.Trend components are long-term non-seasonal patterns in the data.
- **Seasonality :** Any predictable change or pattern in a time series that recurs or repeats over a specific time period can be said to be seasonality.
- **Cyclical :** The cyclical variation in a time series describes the medium-term changes in the series, caused by circumstances, which repeat in cycles. The duration of a cycle extends over longer period of time, usually two or more years.
- **Residual :** These are sudden changes occurring in a time series which are unlikely to be repeated. They are components of a time series which cannot be explained by trends, seasonal or cyclic movements and called as residual or random components.

  Considering the effects of these four components, two different types of models are generally used for a time series viz. Multiplicative and Additive models.
  Multiplicative Model: Y(t) = T(t) × S(t) × C(t) × I(t)
  Additive Model: Y(t) = T(t) + S(t) + C(t) + I(t)

  Here Y(t) is the observation and T(t), S(t), C(t) and I(t) are respectively the trend, seasonal, cyclical and irregular variation at time t.

Multiplicative model is based on the assumption that these four components of a time series are not necessarily independent and they can affect one another; whereas in the additive model it is assumed that these four components are independent of each other.

Note : In our time series data, we didn't find any cyclical component.So, we didn't consider the effect of cyclical component in our model.

5.2 Framework for the used methodology

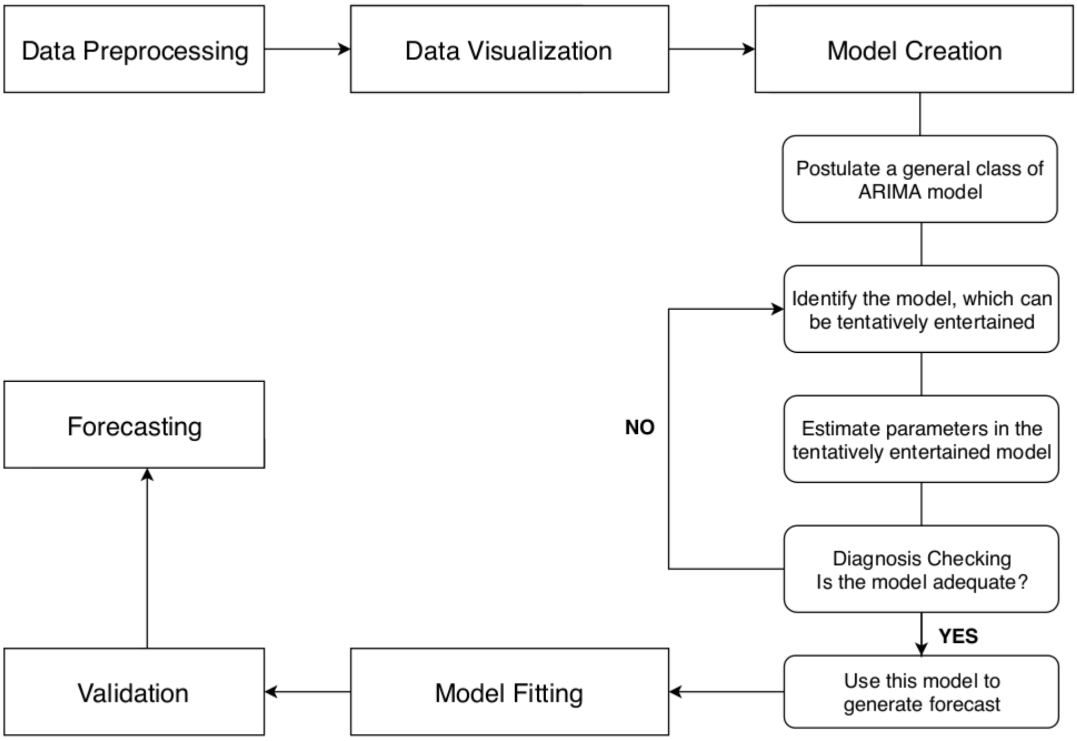

Data Preprocessing
Data Visualization
Model Creation
Postulate a general class of ARIMA model
Identify the model, which can be tentatively entertained
Estimate parameters in the tentatively entertained model
Diagnosis Checking Is the model adequate?
NO
YES
Use this model to generate forecast
Model Fitting
Validation
Forecasting

Figure 1: Framework for the proposed methodologies

5.3 Understanding and pre-processing the data

In our project, we are using Superstore sales data. There are several categories like furniture, office supplies, electronic devices etc in our dataset. Since the time series data is of varying length, we cannot directly build a model on this data set. So how can we decide the ideal length of a series? There are multiple ways in which we can deal with it and here are a few ideas.

- Pre-processing of data includes elimination of columns we do not need, check missing values, aggregate sales by date an-d so on.
- Take the mean of all the lengths(we will take the averages of daily sales value for that month instead), truncate the longer series, and pad the series which are shorter.

Let's suppose we are forecasting for the Furniture sale.

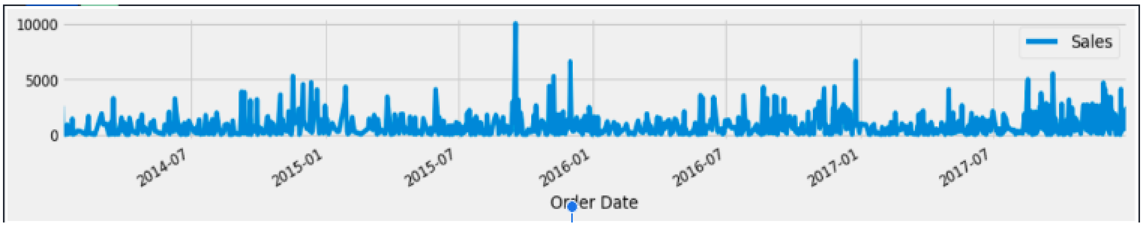


Figure 2: Actual Plot

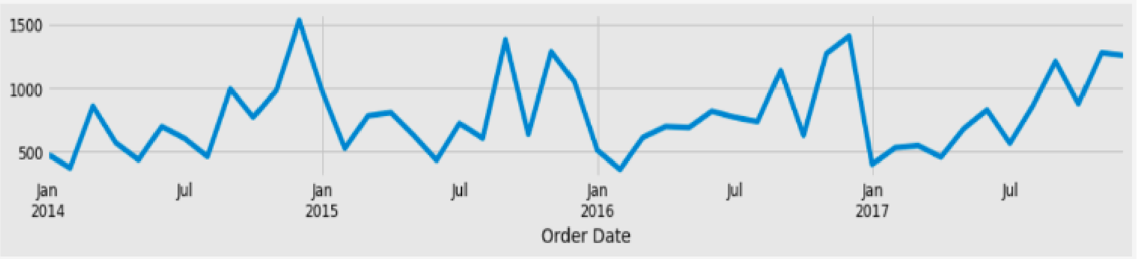


Figure 3: Processed Plot

5.4 Data Visualization

We can visualize our data using a method called time-series decomposition that allows us to decompose our time series data into three distinct components: trend, seasonality, and residual.

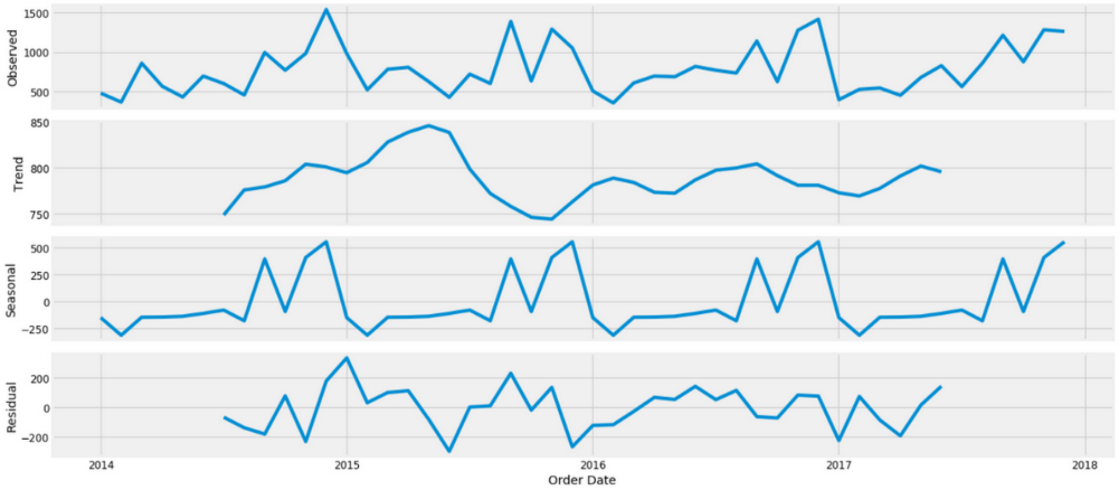


Figure 4: Time series Decomposition

5.5 The ARIMA Time Series Model

One of the most effective methods used in time series forecasting is known as the ARIMA model, which stands for AutoRegression Integrated Moving Average. ARIMA is a model that can be fitted to time series data in order to better understand or predict future points in the series. The idea behind ARIMA is that the model adapts automatically to a given history of data.

ARIMA models are denoted with the notation ARIMA(p,d,q). Together these three parameters account for seasonality, trend, and noise in datasets:

- **p is the auto-regressive** order of the model. It allows us to incorporate the effect of past values into our model. Ex: Y(t) = c + αY(t-1) + e(t), where c is a constant and e(t) is error at time t, and Y(t-1) is value at previous point with a weight. The given equation is a first order lag of AR,denoted as ARIMA(1,0,0).
- **d is the integrated** order of the model. This includes terms in the model that incorporate the amount of differencing (i.e. the number of past time points to subtract from the current value) to apply to the time series. Ex: Y(t) - Y(t-1) = c + e(t), which is denoted as ARIMA(0,1,0) and also known as Random walk.
- **q is the moving average** order of the model. This allows us to set the error

of our model as a linear combination of the error values observed at previous time points in the past. Ex: Y(t) = c + αY(t-1) + βe(t-1) + e(t), where every term has same role as described for p parameter. The extra term e(t-1) is the forecast error term for the first lag multiplied by a coefficient beta. The given equation can be denoted as ARMA(1,1) or ARIMA(1,0,1).

Thus, the order (p,d,q) of the ARIMA model specifies the number of autoregression lags, order of differencing,number of moving average lags, respectively.

5.5.1 Seasonal ARIMA

Seasonality in a time series is a regular pattern of changes that repeats over S time periods, where S defines the number of time periods until the pattern repeats again.

For example, there is seasonality in monthly data for which high values tend always to occur in some particular months and low values tend always to occur in other particular months. When dealing with seasonal effects, we make use of the seasonal ARIMA, also know as SARIMA which is denoted as ARIMA(p,d,q)(P,D,Q)S. Here, (p, d, q) are the non-seasonal parameters described above, while (P, D, Q) follow the same definition but are applied to the seasonal component of the time series. In this case, S = 12 (months per year) is the span of the periodic seasonal behavior. For quarterly data, S = 4 time periods per year The seasonal ARIMA model incorporates both non-seasonal and seasonal factors in a multiplicative model. One shorthand notation for the model is ARIMA(p, d, q) × (P, D, Q)S, with p = non-seasonal AR order, d = non-seasonal differencing, q = non-seasonal MA order, P = seasonal AR order, D = seasonal differencing, Q = seasonal MA order, and S = time span of repeating seasonal pattern.

5.5.2 Concept of Stationarity

The concept of stationarity could be visualized as a form of statistical equilibrium. The statistical properties like mean and variance of a stationary data does not depend upon time. It is a necessary condition for building a time series model that can be used for forecasting future values. It also reduces the mathematical complexity of the fitted model.

5.5.3 Parsimony Principle

According to this principle, always the model with smallest possible number of parameters is to be selected so as to provide an adequate representation of the underlying time series data. One aspect of this principle is that when face with a number of competing and adequate explanations, pick the most simple one. We take the principle into consideration while doing parameter selection.

5.6 Box-Jenkins Methodology

After describing time series models, the next issue to our concern is how to select an appropriate model that can produce accurate forecast based on a description of historical pattern in the data and how to determine the optimal model orders. Statisticians George Box and Gwilym Jenkins developed a practical approach to build ARIMA model, which best fit to a given time series. Their concept has fundamental importance on the area of time series analysis and forecasting. The Box-Jenkins methodology does not assume any particular pattern in the historical data of the series to be forecasted. Rather, it uses a three step iterative approach of model identification, parameter estimation and diagnostic checking to determine the best model from a general class of ARIMA models.. This three-step process is repeated several times until a satisfactory model is finally selected. Then this model can be used for forecasting future values of the time series.

5.6.1 Model Identification and Model Selection:

Making sure that the variables are stationary, and using plots of the autocorrelation and partial autocorrelation functions of the dependent time series to decide which (if any) autoregressive or moving average component should be used in the model.

- **ACF and PACF plots**
  Autocorrelation and partial autocorrelation plots are heavily used in time series analysis and forecasting. These are plots that graphically summarize the strength of a relationship with an observation in a time series with observations at prior time steps. After a time series has been stationarized by differencing, the next step in fitting an ARIMA model is to determine whether AR or MA terms are needed to correct any autocorrelation that remains in the differenced series. By looking at the **autocorrelation function (ACF)** and **partial autocorrelation (PACF)** plots of the differenced series, you can tentatively identify the numbers of AR and/or MA terms that are needed.

You are already familiar with the ACF plot: it is merely a bar chart of the coefficients of correlation between a time series and lags of itself. The PACF plot is a plot of the partial correlation coefficients between the series and lags of itself.

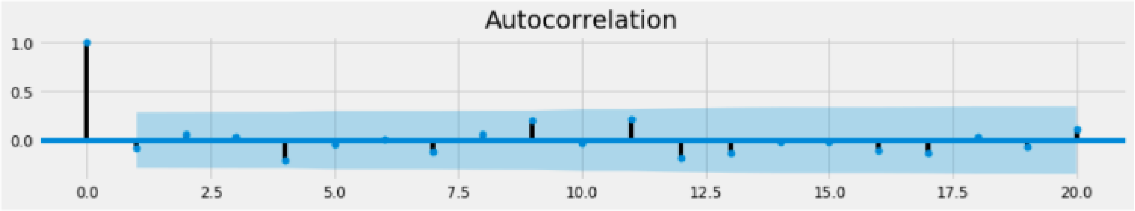


Figure 5: Autocorrelation Plot

5.6.2 Parameter estimation:

When evaluating and comparing statistical models fitted with different parameters, each can be ranked against one another based on how well it fits the data or its ability to accurately predict future data points. We will use the AIC (Akaike Information Criterion) value, which is conveniently returned with ARIMA models fitted using statsmodels. The AIC measures how well a model fits the data while taking into account the overall complexity of the model. A model that fits the data very well while using lots of features will be assigned a larger AIC score than a model that uses fewer features to achieve the same goodness-of-fit. Therefore, we are interested in finding the model that yields the lowest AIC value.

- **Akaike Information Criterion (AIC)**

The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

$$\text{AIC} = \text{2K - 2log}(\widehat{L})$$
$$where, k = number of parameters in the model$$
$$\widehat{L} = maximized likelihood of a model$$

5.6.3 Diagnostic checking:

By testing whether the estimated model confirms to the specifications of a stationary univariate process. In particular, the residuals should be independent of each other and constant in mean and variance over time. (Plotting the mean and variance of residuals over time and plotting autocorrelation and partial autocorrelation of the residuals are helpful to identify misspecification.) If the estimation is inadequate, we have to return to step one and attempt to build a better model.

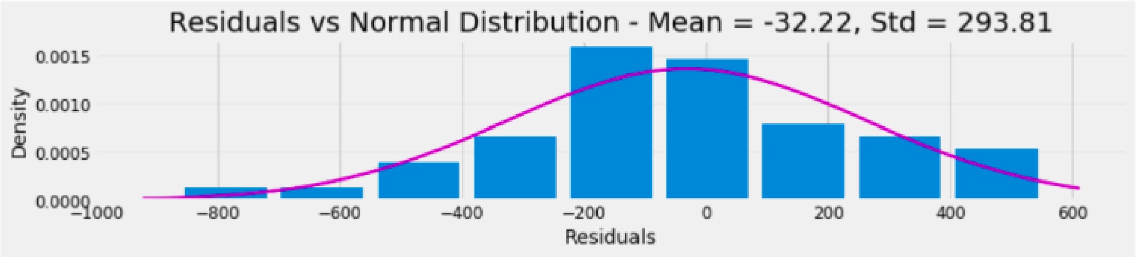
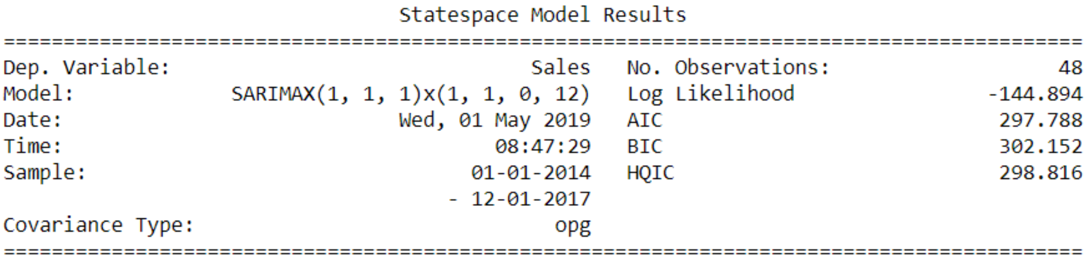


Figure 6: Residual vs Normal Distribution

5.7 Fitting an ARIMA Time Series Model

We have identified the set of parameters that produces the best fitting model to our time series data.



```
                                  Statespace Model Results
==========================================================================================
Dep. Variable:                               Sales   No. Observations:                  48
Model:             SARIMAX(1, 1, 1)x(1, 1, 0, 12)   Log Likelihood                -144.894
Date:                             Wed, 01 May 2019   AIC                            297.788
Time:                                     08:47:29   BIC                            302.152
Sample:                                 01-01-2014   HQIC                           298.816
                                      - 12-01-2017
Covariance Type:                               opg
==========================================================================================
```

Figure 7: Fitted Model Result

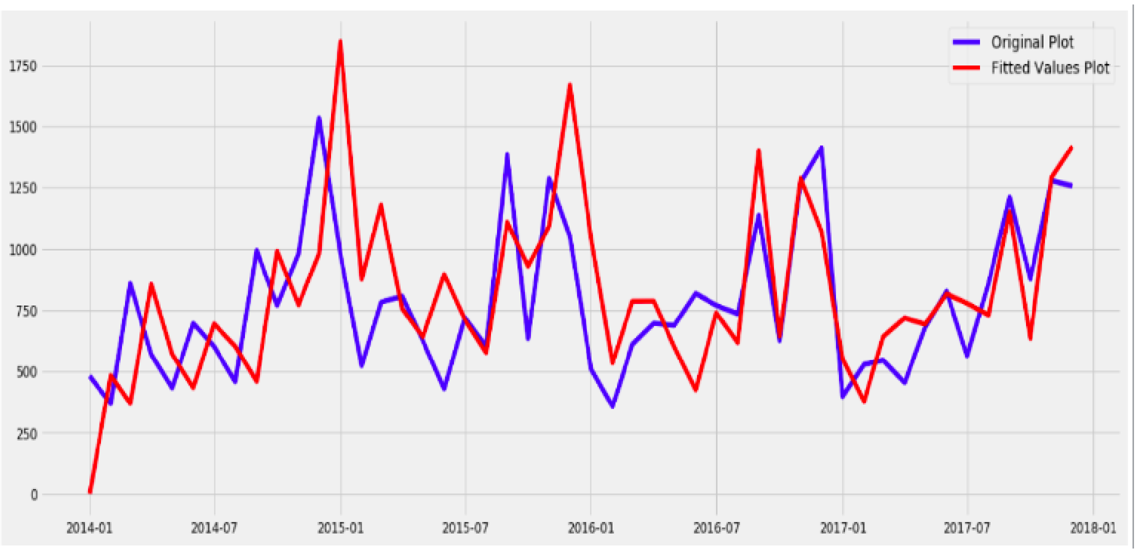


Figure 8: Model Fitting for Furniture Sale

5.8 Validating Forecast

We have obtained a model for our time series that can now be used to produce forecasts. We start by comparing predicted values to real values of the time series, which will help us understand the accuracy of our forecasts.Here, we set our forecasts to start at 2017–01–01 to the end of the data.

The line plot is showing the observed values compared to the rolling forecast predictions. Overall, our forecasts align with the true values very well, showing an upward trend starts from the beginning of the year and captured the seasonality toward the end of the year.

It is also useful to quantify the accuracy of our forecasts. We will use the MSE (Mean Squared Error) and RMSE(Root Mean Squared Error), which summarizes the average error of our forecasts. For each predicted value, we compute its distance to the true value and square the result.

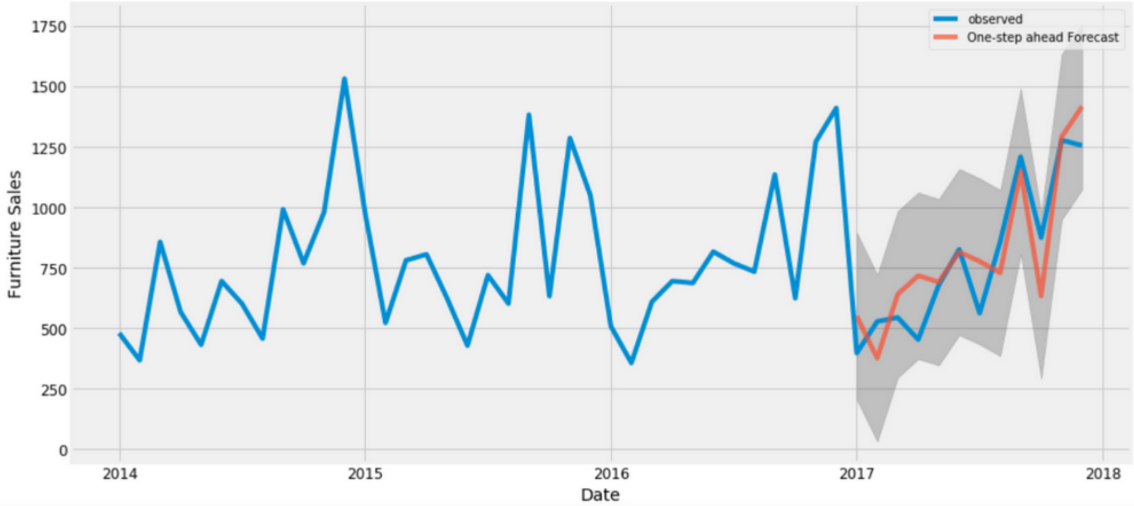


Figure 9: Validating Forecast

We will calculate error by Mean Squared Error formulae :

$$\frac{1}{n}\sum_{i=0}^{i=n}(ForcastedValue\ -ActualValue)^2$$

Figure 10: Mean Squared Error

The Mean Squared Error of our forecasts is 22993.58 and the Root Mean Squared Error of our forecasts is 151.64. Root Mean Square Error (RMSE) tells us that our model was able to forecast the average daily furniture sales in the test set within 151.64 of the real sales. Our furniture daily sales range from around 400 to over 1200.

5.9 Producing and visualizing forecasts

Our model finally captured the furniture sales seasonality.

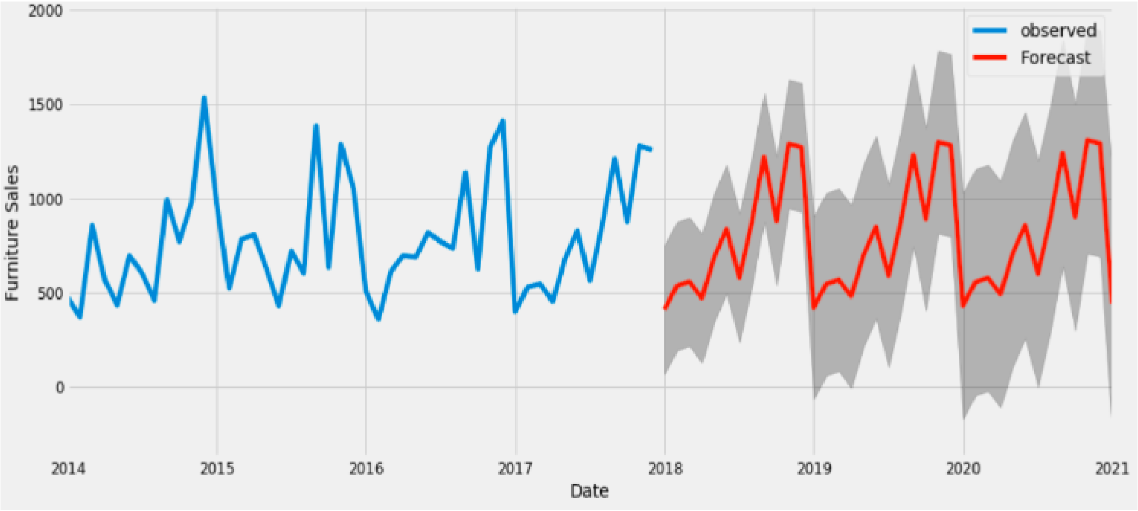


Figure 11: Forecast

6 Resources required

6.1 Hardware Requirements

1. Random Access Memory : 2 GB
2. Intel Core Processing Unit, min., i3 5th Generation.

6.2 Software requirements

1. Python 3
2. Microsoft Excel
3. Cython
4. Pystan
5. C++ Compiler

7 Conclusion

Sales forecasting is an important function of managing supply chain. Its integration with other business functions makes it one of the most important planning processes business can deploy for future.In this context, we developed an ARIMA model for forecasting of the inventory stock for the sales. The results obtained proved that this model can be used for modeling and forecasting the future demand in this product sales. These results will provide the managers with reliable guidelines in making decisions.

8 References

[1] Forecasting of Demand using ARIMA model, Jamal Fattah, Zineb Aman and Haj El Moussami(October 30, 2018)

[2] Sales Prediction With Parameterized Time Series Analysis, Michael Schaidnagel, Christian Abele, Fritz Laux and Ilia Petrov (January 2013)

[3] Time series forecasting using a hybrid ARIMA and neural network model,G. Peter Zhang (January 2003)

[4] Sales Prediction with Time Series Modelling, Gautam Shine and Sanjib Basak (2012)

[5] Time Series Sales Forecasting, James J.Pao and Danielles S.Sullivan (2017)

[6] G.E.P. Box, G. Jenkins, Time Series Analysis, Forecasting and Control, Holden-Day, San Francisco,CA, 1970. (October 30, 1989)

[7] A SURVEY ON ARIMA FORECASTING USING TIME SERIES MODEL,Z. Asha Farhath,B. Arputhamary, Dr. L. Arockiam (August 8, 2016)

[8] An Introductory Study on Time Series Modeling and Forecasting,Ratnadip Adhikari, R. K. Agrawal (Feb 26, 2013))

[9] Original source: P. J. Brockwell, R. A. Davis, “Introduction to Time Series and Forecasting”, 2nd edition, Springer Publication; March, 2003.